



3 1761 11633877 3

Government
Publications

62-Z-503

Statistics Canada.

Data quality and record linkage -
an experiment.



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761116338773>



Statistics
Canada

Statistique
Canada

62-Z-503

**Consumer Income
and Expenditure
Division**

**Division du revenu et
des dépenses
des consommateurs**

Data quality and
record linkage -
An experiment

Qualité des données et
couplage des dossiers -
Essai

LIBRARY

2175



Statistics Canada

Statistique Canada

Government
Publications

Consumer Income and
Expenditure Division

Division du revenu et des
dépenses des consommateurs

Research Section

Section de la recherche

Data quality and
record linkage -
An experiment

Qualité des données et
couplage des dossiers -
Essai



Published under the
authority of the
Minister of Supply
and Services Canada

Publication autorisée
par le ministre des
Approvisionnements
et Services Canada

Statistics Canada
should be credited
when reproducing or
quoting any part
of this document

Reproduction ou citation
autorisée sous réserve
d'indication de la source:
Statistique Canada

©Minister of Supply and
Services Canada 1982

© Ministre des Approvision-
nements et Services Canada 1982

February 1982
8-3303-520

Février 1982
8-3303-520

Ottawa

Ottawa

SYMBOLS

The following standard symbols are used in Statistics Canada publications:

.. figures not available.

... figures not appropriate or not applicable.

- nil or zero.

-- amount too small to be expressed.

P preliminary figures.

r revised figures.

x confidential to meet secrecy requirements of the Statistics Act.

SIGNES CONVENTIONNELS

Les signes conventionnels suivants sont employés uniformément dans les publications de Statistique Canada:

.. nombres indisponibles.

... n'ayant pas lieu de figurer.

- néant ou zéro.

-- nombres infimes.

P nombres provisoires.

r nombres rectifiés.

x confidentiel en vertu des dispositions de la Loi sur la statistique relatives au secret.

REFAÇE

Record linkage is a technique of enriching the information value of data files. This technique combines data from diverse sources and has been made possible through technological advancements in electronic data processing methods. Apart from enriched data files, improved data quality, reduced response burden, and enhanced capabilities to judge data adequacy are noteworthy side effects of record linkage.

Statistics Canada must endeavour to avail itself of modern cost-effective techniques. To this end, the experiment described in this report was undertaken. While theoretical reference material was readily available, relevant practical examples were scarce. We made extensive use of the experience gained by members of the United States Department of Health, Education and Welfare in their development of linkage applications.

This report puts the emphasis on describing the techniques used in the practical context of two concrete data files, the problems encountered and how they were dealt with, and assesses the quality of income reporting in the 1971 Census compared to 1971 tax data. As such, it is addressed primarily to personnel involved in record linkage studies and users of census income data.

The Bureau's linkage experiment was carried out by members of the research staff of the Consumer Income and Expenditure Division of Statistics Canada under the direction of Mrs. G. Oja. All computer-related activities were directed by L. Nolet, and H.E. Alter was responsible for the specification of matching routines, data preparation and analysis. He will also deal with technically oriented inquiries resulting from this study.

PRÉFACE

Le couplage des dossiers accroît la valeur informative des fichiers de données. La technique consiste à grouper des données de diverses sources et a été rendue possible par les progrès réalisés dans le domaine des méthodes de traitement électronique des données. Le couplage des dossiers nous permet en outre d'améliorer la qualité des données, de réduire le fardeau des répondants et de mieux juger de la pertinence des données.

Statistique Canada doit s'efforcer de mettre à profit les techniques modernes de réduction des coûts. C'est dans cette perspective que l'expérience décrite ici a été entreprise. Bien que les documents de référence théoriques en cette matière soient facilement accessibles, les exemples concrets manquent. Nous nous sommes donc largement appuyés sur l'expérience acquise par le ministère américain de la Santé, de l'Éducation et du Bien-être dans ses travaux de couplage.

Le présent rapport s'emploie à décrire les techniques utilisées dans le contexte pratique de deux fichiers de données concrets et à présenter les problèmes qui ont surgi et les solutions qu'on a retenues; en outre, il compare la qualité des chiffres sur le revenu tirés du recensement de 1971 avec les données fiscales de la même année. Il s'agit donc d'un ouvrage qui intéressera principalement les individus qui exécutent des études sur le couplage des dossiers et les utilisateurs des données du recensement qui portent sur le revenu.

L'expérience de couplage du Bureau a été faite par des membres de l'équipe de recherche de la Division du revenu et des dépenses des consommateurs de Statistique Canada, sous la direction de G. Oja. L. Nolet était responsable de toutes les activités à caractère informatique tandis que H.E. Alter devrait s'occuper de la définition des routines d'appariement, de la préparation et de l'analyse des données. Ce dernier s'occupera aussi des demandes de nature technique que pourra susciter la présente étude.

TABLE OF CONTENTS

	Page
Summary	9
Record Linkage	11
Introductory Remarks	11
The Need for Linked Data	14
The Matching of Tax and Census Records	15
Post-match Analysis	21
Methodological Review	21
Income Reporting on Matches and Non-matches	29
True Matches and Reporting Errors	31
Income Composition	33
Non-matched Records from the 1971 Census	46
Postscript	50

Text Table

I. Matches by Time of Occurrence of Decision Type in Census-RC-T Match, 1971	21
II. Accuracy Levels and Computer Time Levels for the Census-RC-T Match, 1971	22
III. Address Components by Consistency Status and Agreement Type for the Census-RC-T Match, 1971	25
IV. Personal Characteristics and Variables by Consistency Status and Agreement Type for the Census-RC-T Match, 1971	27

Table

1. Census Income Recipients, by Match Status and by Major Source of Income for Income Base Year, 1970	56
---	----

TABLE DES MATIÈRES

	Page
Résumé	9
Couplage des dossiers	11
Remarques liminaires	11
Utilité des données couplées	14
Appariement des dossiers de l'impôt et du recensement	15
Analyse post-appariements	21
Méthodologie	21
Déclaration du revenu - Dossiers appariés et non appariés	29
Appariements justes et erreurs de déclaration	31
Composition du revenu	33
Dossiers non appariés du recensement de 1971	46
Post-scriptum	50

Tableau explicatif

I. Appariements selon le moment du type de décision, appariement recensement-RC-I, 1971	21
II. Niveau de précision et temps d'ordinateur, appariement recensement-RC-I, 1971	22
III. Éléments de l'adresse selon le statut de cohérence et le type de convergence, appariement recensement-RC-I, 1971	25
IV. Caractéristiques et variables personnelles selon le statut de cohérence et le type de convergence, appariement recensement-RC-I, 1971	27

Tableau

1. Personnes ayant déclaré un revenu au recensement, selon le statut d'appariement et la principale source de revenu, 1970	56
--	----

TABLE OF CONTENTS - Continued

Table	Page
2. Adult Census Population by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971	56
3. Sample of Adult Census Population by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971	57
4. Matched Records by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971	57
5. True Matches with Income Reported in One Source Only, by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971	58
6. All Non-matches by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971	58
7. True Non-matches by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971	59
8. False Non-matches by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971	59
9. Reporting Pattern of Income Components for True Matches, by Source Presence with Consistency Rate for Income Base Year, 1970	60
10. Reporting Patterns of Income Components, by Province for True Matches with Income Subject to Taxation Present in Both Sources for the Income Base Year, 1970	61
11. Omissions and Substitutions of Income Components, by Reliability Category with Aggregate Total Income by Data Source for the Income Base Year, 1970	62
12. Income Effect of Component Omission by Reliability Category, by Source of Omission, by Incidence Group for Income Base Year, 1970	64
13. Income Effect of Component Substitution, by Reliability Category and Incidence of Substitution for Income Base Year, 1970	65

TABLE DES MATIÈRES - suite

Tableau	Page
2. Population adulte du recensement par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971	56
3. Échantillon de la population adulte du recensement par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971	57
4. Dossiers appariés par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971	57
5. Appariements justes, revenu déclaré dans une seule source par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971	58
6. Ensemble des non-appariements par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971	58
7. Non-appariements justes par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971	59
8. Non-appariements erronés par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971	59
9. Modèle de déclaration des éléments du revenu dans les appariements justes, selon la présence de la source et taux de cohérence, 1970	60
10. Régime de déclaration des éléments du revenu, par province, appariements justes, revenu soumis à l'impôt présent dans les deux sources, 1970	61
11. Omissions et substitutions d'éléments du revenu, par catégorie de fiabilité, et revenu agrégatif total par source de données, 1970	62
12. Effet sur le revenu de l'omission d'un élément, par catégorie de fiabilité, selon la source de l'omission et le groupe d'incidence, 1970	64
13. Effet sur le revenu de la substitution d'un élément, par catégorie de fiabilité, selon la source de la substitution et le groupe d'incidence, 1970	65

TABLE OF CONTENTS - Concluded

Table	Page
1. Income Effect of Combined Omissions and Substitutions of Income Components, by Reliability Category and Source of Omissions for Income Base Year, 1971	66
2. Match Rates and Taxfiler Rates with Components, by Province with Descending Rank Order for the Income Base Year, 1970	67
3. Substitution of Employment Income Components for the Income Base Year, 1970	68
4. Census Gains and Losses Vis-à-vis RC-T Reporting as a Result of Component Substitution for the Income Base Year, 1970	68
5. Number and Percentage of Records Appearing in Equivalent and Neighbouring Income Classes for Selected Income Components whenever Component has been Reported in at Least One Source (Census or RC-T) for the Income Base Year, 1970	69
6. Distribution of Income Subject to Taxation with Class Deficiency Rates, by Match Status for the Income Base Year, 1970	70
7. Provincial Rank Order of Match Rates and Success Rates with Supporting Data for the Income Base Year, 1970	71
8. Bibliography	73

TABLE DES MATIÈRES - fin

Tableau	Page
14. Effet sur le revenu des omissions et des substitutions des éléments du revenu, par catégorie de fiabilité et source des omissions, 1971	66
15. Éléments des taux d'appariement et des taux de déclaration à l'impôt, par province et par ordre décroissant, 1970	67
16. Substitution d'éléments du revenu de l'emploi, 1970	68
17. Gains et pertes du recensement par rapport à RC-I résultant de la substitution d'éléments, 1970	68
18. Nombre et pourcentage de dossiers paraissant dans des classes de revenu équivalentes ou voisines en fonction de certains éléments du revenu déclarés dans au moins une source (recensement ou RC-I), 1970	69
19. Répartition du revenu soumis à l'impôt et déficit, par statut d'appariement, 1970	70
20. Classement des taux d'appariement et des taux de réussite par province, 1970	71
Bibliographie	73

This report presents the results of an experiment which makes use of up-to-date computer technology and methodological developments with respect to record linkage. The data bases used, a sample of 1971 Census records, and tax records for the 1970 reference year, were the most useful data available when the project was started, but are out of date at the time of publication. The study must be seen in its usefulness of opening up new methods and in suggesting technical and cost feasibility of similar projects. Analytical conclusions concerning income reporting patterns and other data problems may have to be viewed with reservations in the light of current collection, edit and imputation practices, which may depart from those used a decade ago.

The introductory section of this report discusses aspects of information value with reference to data and the possible improvement of the information value through record linkage. Types of record linkage, namely statistical and exact matches, are discussed next, and the exploitation of "administrative" data is advocated.

The need for linked data is demonstrated in the light of perceived utility of longitudinal records. Such records could be constructed as a follow-up to the initial linkage. Other benefits, such as an evaluation of the data quality of census records, would be obtained as a side benefit.

The matching process and the underlying conceptual and technical problems are discussed in sufficient detail to provide some stimulation for a professional audience. The informed layman may find it too detailed and hence boring. The reasons for selecting a sample from the Census rather than from the tax universe, the choice of matching variables, and the execution of matching routines all fall into this section.

The post-match analysis distinguishes between a methodological review, which assesses the efficacy of the linkage operation, and an assessment of income reporting. The assessment of income reporting in the Census, while it can only observe differences from two sources for identical respondents, actually must cope with the combined effect of non-response or partial response, component substitution due to misinterpretation or instrumental differences (tax form, census questionnaire), and processing errors including differences caused by deliberate acts of edit and imputation.

On trouvera ici les résultats d'un essai qui s'appuie sur les techniques informatiques et les méthodes les plus modernes dans le domaine du couplage des dossiers. Les bases de données utilisées, c.-à-d., un échantillon de dossiers du recensement de 1970 et les déclarations d'impôt pour l'année de référence de 1970, étaient les meilleures données accessibles au début du projet. Toutefois, elles sont maintenant dépassées. La présente étude est donc surtout utile parce qu'elle ouvre la voie à de nouvelles méthodes et qu'elle donne une idée de la faisabilité technique et de la rentabilité de travaux similaires. Comme les pratiques actuelles de collecte, de contrôle et d'imputation peuvent différer de celles qui étaient en usage il y a 10 ans, les structures de déclaration et autres problèmes observés doivent être examinés avec quelques réserves.

Dans les remarques liminaires, on discute de la valeur informative associée aux données et de son amélioration éventuelle au moyen du couplage des dossiers. On traite ensuite des différents types de couplage, c.-à-d., les appariements statistiques et exacts, et on recommande l'exploitation des données administratives.

L'utilité des données couplées réside dans la valeur perçue des dossiers longitudinaux qui pourraient être construits à la suite du couplage initial. Ces dossiers offrent aussi d'autres avantages, tels que l'évaluation de la qualité des données tirées des dossiers du recensement.

Le processus d'appariement et les problèmes conceptuels et techniques sous-jacents font l'objet d'une discussion détaillée qui intéressera les spécialistes. Aux yeux du non-spécialiste bien informé, la discussion sera peut-être trop détaillée et par conséquent fastidieuse. Cette partie explique les raisons pour lesquelles on a prélevé l'échantillon des dossiers du recensement plutôt que de l'univers des déclarations d'impôt, le choix des variables d'appariement et l'exécution des routines d'appariement.

L'analyse post-appariement fait la distinction entre la revue méthodologique, qui évalue l'efficacité du couplage, et l'évaluation de la déclaration des revenus. Bien que cette dernière évaluation ne s'appuie, dans le cas du recensement, que sur les différences observées entre deux sources pour des répondants identiques, elle doit tenir compte des effets combinés des non-réponses et des réponses partielles, de la substitution d'éléments à la suite de différences dans l'interprétation et l'instrument (formulaire de l'impôt, questionnaire du recensement) et des erreurs de traitement, y compris les écarts voulus attribuables au contrôle et à l'imputation.

The effect of reporting differences is presented in various dimensions. It will be of greater interest to the producer of statistics rather than the user. Some of the users may view the observed differences with suspicion, but they should be reminded that data and statistics contain imperfections just like any other product. The producers' endeavour to control and improve data quality is reflected in the undertaking of studies such as this one.

Non-matches form a subset of particular interest. They do not contribute anything to the assessment of data quality, but they do invite a number of questions. Non-matches are largely the result of missing counterparts of census records in the tax universe. Such absences are justified as a rule, but they underline the limitation of statistical information derived solely from tax records.

The **postscript** contains conclusions of a fairly general nature; i.e., not only as far as this particular exercise is concerned, but concerning data collection and quality control in general wherever such data are to be used for linkage applications. The postscript also points towards alternative linkage applications or alternative data sources for such application. These alternatives are not presented with the firmness of a recommendation, but they could provide a basis for discussion.

Les effets des différences de déclaration sont présentés sous plusieurs angles et intéresseront plutôt les producteurs de statistiques que leurs utilisateurs. Certains utilisateurs éprouveront de la méfiance envers les différences observées mais ils ne doivent pas oublier que les données et les statistiques, comme tout autre produit, ne sont pas parfaites. La réalisation d'études comme celle-ci traduit les efforts de producteurs en vue de contrôler et d'améliorer la qualité des données.

Les **non-appariements** forment un sous-ensemble d'un intérêt particulier. Ils ne contribuent rien à l'évaluation de la qualité mais ils suscitent un certain nombre de questions. Les non-appariements résultent principalement de l'absence dans l'univers des déclarations d'impôt d'éléments correspondants des dossiers du recensement. Ces absences sont généralement justifiées mais elle font ressortir les limites des informations statistiques tirées uniquement de déclarations d'impôt.

Le **post-scriptum** renferme des conclusions d'une portée générale qui vont au-delà de la présente étude et concernent la collecte et le contrôle de la qualité des données en général dans tous les cas où celles-ci servent au couplage. Le post-scriptum propose aussi d'autres applications du couplage ou sources de données utiles dans ce contexte. Celles-ci n'y sont pas présentées comme des recommandations mais comme une base possible de discussion.

Introductory Remarks

Data and **information** are not necessarily synonymous. An abundance of unrelated data may convey little information, whereas a small but well-selected data base may demonstrate high informative values. The degree of usefulness derived therefrom, of course, depends on the needs of the analyst.

One of the great challenges facing statistical agencies is to meet the increasing data needs of users involved in monitoring and assessing current economic or social policies, and developing alternative policy proposals. Business users need information to improve their productivity, academics strive for a better understanding of contemporary phenomena, and the general public wishes to be better informed about its environment. To satisfy these demands for information with a minimum of expenditure and without imposing undue response burden on the population, **record linkage** offers itself as an alternative to conventional data collection and data assembly methods.

The word **linkage** implies that records from existing data sources are combined, thereby forming a data base which contains more comprehensive information than its constituent parts.

While such a procedure may have cost advantages for the Bureau, and while respondents are spared the answering of questionnaires, there are also some negative aspects to be considered. These are caused by the public's perception of record linkage activities.

Statistics Canada is aware of the general feeling of concern regarding record linkage. It is important, however, to keep in mind a basic distinction between different record linkage applications. This distinction relates to the intended end use of the linked files and the extent to which their confidentiality is protected. At least two basic categories of end use must be distinguished: administrative and statistical. Use of data for **administrative purposes** implies that information about an identifiable individual is directly used in some decision process which relates to the individual. Most of the concern about record linkage applies to this type of use, since individuals have no control over someone's combining information about them; information which they supplied to different data collection agencies for different purposes.

Remarques liminaires

Les concepts associés aux termes **données** et **information** ne sont pas nécessairement synonymes. Des masses de données disparates peuvent avoir un contenu informatif négligeable; en revanche, une base de données peu étendue, mais bien choisie peut avoir une valeur informative élevée. L'utilité des données se mesure donc en fonction des besoins de l'analyste.

L'un des grands défis que doivent relever les organismes statistiques est celui de satisfaire aux besoins en données croissants d'utilisateurs qui sont chargés de contrôler et d'évaluer les politiques économiques ou sociales actuelles et d'élaborer des politiques de rechange. En outre, les entreprises ont besoin d'information pour améliorer leur productivité, les universitaires cherchent à mieux cerner les phénomènes contemporains et le grand public désire en savoir davantage sur son environnement. Pour répondre à toutes ces demandes de renseignements en minimisant les frais et le fardeau de réponse imposé aux enquêtés, il conviendrait peut-être de recourir au **couplage des dossiers** plutôt qu'aux méthodes conventionnelles de collecte et de rassemblement des données.

Le terme **couplage** indique que les dossiers émanant de sources existantes de données sont groupés en vue de constituer une base de données qui aura une plus grande valeur informative que ses éléments constitutifs.

Bien que cette méthode soit rentable pour le Bureau et qu'elle dispense les enquêtes de l'obligation de remplir des questionnaires, elle comporte néanmoins certains aspects négatifs liés à la perception qu'a le public des activités entourant le couplage des dossiers.

Statistique Canada est bien conscient des préoccupations qui se rattachent au couplage des dossiers. Il importe toutefois de rappeler une distinction fondamentale entre les diverses applications du couplage, distinction qui met en cause l'utilisation finale des dossiers couplés et la mesure dans laquelle on en protège le secret. En effet, il existe au moins deux grandes catégories d'utilisation finale, à savoir l'administrative et la statistique. Dans le cas d'une utilisation de données à **des fins administratives**, l'information concernant un individu identifiable est directement consacrée à un processus décisionnel qui affecte l'individu. La plupart des critiques formulées contre le couplage des dossiers visent effectivement ce genre d'utilisation, puisque les particuliers n'ont ainsi aucun contrôle sur celui qui recueille à leur sujet des renseignements qu'ils ont fournis pour divers motifs à différents organismes de collecte. Le

Using these combined data in an administrative or decision-making context directly affects the individual, perhaps in a fashion which he or she may consider harmful. By contrast, when linked data are used for a statistical purpose, the resultant file is utilized only to provide statistical aggregates or distributions, while keeping the identity of the individuals concerned strictly confidential. In the case of Statistics Canada, such confidentiality is guaranteed by legislation which contains severe penalties against its violation. Furthermore, over and above the legal obligations, in the present application extraordinary care has been taken to ensure that the required confidentiality of the data is actually preserved in practice.

The choice of the data sources for linkage is critical. Data to be used as linkage characteristics have to be conceptually compatible, their reference periods must be identical, or must be capable of being made to conform by some adjustment process. Furthermore, data used for linkage decisions should be of exceptional quality.

These characteristics of input data to linkage processing are not exhaustive, but they are important for the understanding of the process. The linking of data is not extraordinarily difficult, but whether such an augmented data base contains simply more data or whether it contains more information depends largely on the choice and quality of the input data.

One must distinguish between **synthetic or statistical links** and **direct or exact links**. A synthetically linked record combines data from two or more records, where these combined data refer typically to different units having similar characteristics. One of the pioneering efforts can be attributed to Okner.(1) In the Canadian context a linkage of two household surveys as described by Alter(2) serves to illustrate this type of data assembly.

Direct linkage combines data originating with the same identifiable unit, such as the individual, the family, or the corporation. One form of direct linkage employs records of individuals, where these records originate in different time periods, although they had been submitted to the same agency; e.g., the linkage of tax returns by Revenue Canada to previous returns of the same individual for averaging purposes.

See footnote(s) at end of text.

recours à des données regroupées dans un contexte d'administration ou de prise de décision touchent directement l'individu d'une manière qu'il peut juger comme nuisible. Par contre, lorsqu'on sert de données couplées à des fins statistiques, le dossier qui en résulte n'est utilisé que pour présenter des agrégations statistiques ou des distributions tout en gardant l'identité de l'individu strictement confidentielle. En ce qui concerne Statistique Canada, la confidentialité est assurée par une loi dont la violation est assortie de peines sévères. Outre cette protection officielle, toutes les mesures ont été prises pour que soit effectivement respectée en pratique la confidentialité des données.

Le choix des sources de données à coupler est de prime importance. Les données qui serviront de caractéristiques de couplage doivent être conceptuellement compatibles, avoir la même période de référence ou, à défaut, pouvoir subir les ajustements requis. Enfin, les données déterminant les décisions de couplage devraient être d'une qualité exceptionnelle.

Cette liste de caractéristiques des données soumises au couplage n'est pas exhaustive, mais elle facilitera la compréhension du processus. Le couplage proprement dit ne pose pas de problèmes extraordinaires; la réussite de l'opération tient largement au choix et à la qualité des données d'entrée: la nouvelle base de données ne contient pas simplement plus de renseignements, sa valeur informative s'est accrue.

Établissons tout d'abord une distinction entre les couplages **synthétiques ou statistiques** et les couplages **directs ou exacts**. Le couplage synthétique groupe des données tirées de deux dossiers ou plus portant sur des unités distinctes qui ont des caractéristiques analogues. Okner(1) a joué en cette matière un rôle de pionnier. Au Canada, le couplage de deux enquêtes-ménage décrit par Alter(2) illustre bien ce genre d'opération.

Dans le couplage direct, on groupe des données qui émanent d'une même unité: une personne, une famille, une entreprise. Dans l'une des formes de couplage direct, on groupe des déclarations portant sur des périodes différentes, mais soumis par le même organisme; c'est ce qu'on fait Revenu Canada, par exemple, quand on groupe les déclarations d'impôt d'une même personne pour établir des moyennes.

Voir note(s) à la fin du texte.

Direct linkage for statistical applications usually involves the identification of units of observation in various data sources and subsequently the combining of these records. Like linking parts of a jig-saw puzzle, the completed picture will be more informative than the impressions gained from its disjointed parts.

While the unit of observation has to be identified without ambiguity in order to facilitate the link, the identity has no statistical value and is removed once data assembly has been completed. A parallel exists in standard survey methodology, where identities are known for control purposes and follow-up procedures, but where identities do not form part of statistical working files.

Moreover, recognizing the basic nature of administrative as opposed to statistical uses of data, and given the strict confidentiality provision of the Statistics Act, this Act explicitly authorizes access by Statistics Canada for statistical purposes to Revenue Canada-Taxation (RC-T) tax files. Such access is carefully controlled in agreement with RC-T and, of course, most emphatically does not involve any access by RC-T to individually identifiable data held by Statistics Canada.

In completing the general overview it is worth noting that the most comprehensive work in creating linked data files was probably carried out by the Social Security Administration of the United States Department of Health, Education and Welfare, in cooperation with the United States Bureau of the Census and the Internal Revenue Service.(3) In Canada, work of considerable practical value was done under the auspices of Atomic Energy of Canada Limited.(4) Work of appreciable theoretical interest in the context of direct linkage is contained in an article by Fellegi and Sunter.(5)

Some of the work cited above served as a guide or as inspiration. Some findings herein confirmed that theory has to be tempered with empiricism in order to tailor imperfect data to an operationally feasible methodology.

With these introductory remarks in mind, the rationale for performing a direct link of a sample of 1971 Census records and of compatible tax returns will now be explained. The rationale is justified by data needs, but these could not be satisfied without modern computer technology.

Dans ses applications statistiques, le couplage direct comporte généralement l'identification d'unités d'observation de diverses sources de données, puis le couplage des dossiers voulus. Comme dans un casse-tête, l'image finale nous renseigne plus que l'impression dégageée par chacune des pièces.

Bien que l'unité d'observation doive être définie sans ambiguïté pour faciliter le couplage, son identité n'a aucune valeur statistique, et on n'en tient pas compte une fois que le couplage des données est terminé. On peut faire à cet égard un parallèle entre le couplage et la méthodologie de la majorité des enquêtes; en effet, les moyens d'identification utilisés à des fins de contrôle et de suivi ne font pas partie des fichiers de travail statistiques.

Il existe, rappelons-le, une nette distinction entre les utilisations administratives des données et leurs applications statistiques; de plus, la Loi sur la statistique renferme des dispositions rigoureuses en matière de confidentialité et autorise explicitement, à des fins statistiques, l'accès de Statistique Canada aux dossiers fiscaux de Revenu Canada-impôt (RC-I). L'accès est soigneusement contrôlé en collaboration avec RC-I, mais ce ministère n'est évidemment pas autorisé à consulter, de quelque manière que ce soit, les données identifiables de Statistique Canada.

Avant de mettre un terme à cette introduction, il convient de noter que le plus important travail de couplage a probablement été réalisé par l'Administration de la sécurité sociale du ministère américain de la Santé, de l'Éducation et du Bien-être, en collaboration avec le bureau américain du recensement et le service du revenu(3). Au Canada, des travaux d'une valeur pratique indéniable, ont été faits sous les auspices d'Énergie atomique du Canada Limitée(4). Enfin, on trouvera un présentation de travaux de couplage direct ayant un intérêt théorique manifeste dans un article de Fellegi et Sunter(5).

Certains de ces travaux nous ont guidés ou inspirés. Les résultats qu'on y présente confirment que la théorie doit s'appuyer sur l'expérience si l'on veut pouvoir adapter des données imparfaites à une méthodologie opérationnelle.

Ces remarques étant faites, nous verrons maintenant pourquoi nous avons tenté de coupler directement un échantillon de dossiers du recensement de 1971 à un groupe compatible de déclarations d'impôt. À la base, l'opération était justifiée par un besoin de données que seule la technique informatique moderne pouvait satisfaire.

Voir note(s) à la fin du texte.

The Need for Linked Data

Need can best be expressed in terms of more information for the purpose of analysing complex issues. A link of census data and tax records provides more information by presenting a more complete picture of the population than any one of these sources does in isolation. Cost and technical limitations, however, prohibit the linking of all census records to tax records. Thus, a sample of census data had to be employed, and a **sample of Census data** is all that can feasibly be used, given present technology available to the Bureau and resulting resource constraints.

The linked file shows a number of advantages over its constituent parts. It combines detailed socio-demographic data from the Census with reliable income data from Revenue Canada-Taxation (RC-T). Remember that RC-T records are virtually devoid of socio-demographic data, whereas Census income information is probably subject to appreciable reporting errors. The linked file thus permits a study of the quality of income reporting in the Census. It also permits the quantification and analysis of errors of income reporting in the Census.(6) Furthermore, a linked file permits income aggregation for families on an RC-T basis, despite the RC-T impediment of identifying individuals without associating them with family units. Census individuals, on the other hand, can be placed as members of family units. Consequently, established census family relationships in a linked file will supply family income as reported to RC-T.

The most important reason for creating a linked file must be seen in its capability to be updated annually with tax records for individuals in the sample, provided they remain taxfilers. The resulting data base permits longitudinal analysis, i.e., the analysis of identical units over time. Usually, longitudinal data have to be collected through repeated interviews by way of **panel surveys**.(7)

Longitudinal data satisfy a number of needs which cannot be met by cross-section data. Any single set of cross-section data, of course, is only a snapshot of the universe at a specific point in time. To measure change, the use of several cross-section series is required, but such a trend

See footnote(s) at end of text.

Utilité des données couplées

L'utilité des données couplées tient au fait que leur grande valeur informative facilite l'analyse de questions complexes. Le couplage de données du recensement et des déclarations d'impôt, par exemple, nous donne une meilleure image de la population que chacune de ces sources de renseignements prise isolément. Des contraintes d'ordre financier et technique rendant impossible le couplage de tous les dossiers du recensement à des déclarations d'impôt, nous avons dû utiliser un échantillon de données du recensement. D'ailleurs, seul un **échantillon de données** peut raisonnablement faire l'objet d'un couplage vu les possibilités techniques dont dispose aujourd'hui le Bureau et les compressions budgétaires que cela entraîne.

Les dossiers couplés offrent de nombreux avantages. Ils nous permettent de coupler les données socio-démographiques détaillées du recensement aux chiffres fiables sur le revenu de Revenu Canada-impôt (RC-I). On se rappellera que les dossiers de RC-I ne contiennent pratiquement aucune donnée socio-démographique, alors que les chiffres du recensement sur le revenu sont probablement entachés d'erreurs de déclaration appréciables. Le couplage des dossiers nous permet donc d'étudier la qualité de la déclaration de revenu à l'occasion du recensement. Grâce au couplage, on peut également quantifier et analyser les erreurs de déclaration des revenus lors du recensement(6). Enfin, l'opération rend possible le calcul du revenu familial à partir des chiffres de RC-I, bien que Revenu Canada, par souci d'efficacité, n'associe pas les personnes aux unités familiales auxquelles elles appartiennent. Au recensement, par contre, les individus peuvent être rattachés à une unité familiale. Le couplage nous permet donc de calculer les revenus familiaux en fonction des chiffres déclarés à RC-I.

Le principal facteur motivant la création d'un dossier couplé est l'aptitude de celui-ci à être mis à jour chaque année grâce aux dossiers financiers des individus qui composent l'échantillon, condition qu'ils continuent de produire la déclaration. La base de données résultante favorise l'analyse longitudinale, c'est-à-dire l'analyse d'unités identiques dans le temps. Les données longitudinales sont normalement recueillies au moyen d'interviews successifs auprès d'un **échantillon constant**(7).

Les données longitudinales répondent à certain nombre de besoins que ne peuvent satisfaire les données transversales. Un seul ensemble de données transversales ne représente, bien sûr, qu'un instantané de l'univers à un moment donné. Afin de mesurer la variation, il faut faire appel à plusieurs séries transversales; l'analyse

Voir note(s) à la fin du texte.

analysis can only measure **net change**, not **gross change**. Moreover, causality is difficult to infer from trend analysis. The causal relationships, if they exist, and if they are time-related, can be more easily studied with longitudinal data. It is here, that panel surveys would have an advantage over "updating procedures" because special questions to probe or establish causality could be asked.

The understanding of causality is very important for policy formulation and the monitoring of policy impact. Policy may be designed to produce change, or it may be designed to be neutral, but the intent may or may not coincide with subsequent events. To be able to measure the resulting policy effects the time frame as well as the gross-change concept are critical. In the extreme, a zero net change may be interpreted as a neutral policy effect, whereas in reality, there was a positive effect on one part and a negative effect on another part of the population, having resulted in "transfers" - an effect which is quite different from a zero impact.

There are technical and cost considerations that have to be evaluated when contemplating the inception and use of longitudinal data. However, this decision is at least one step removed from the present undertaking because linkage is a necessary condition for producing a longitudinal file. Consequently, the task at hand had to do with the feasibility and cost effectiveness of record linkage, although long-range objectives had to be kept in mind.

Matching of Tax and Census Records

The cost and the technical feasibility of census-taxation link had been explored in a pilot study. This pilot study contained slightly over 2,000 households in Eastern Ontario. The study also contributed appreciably towards the development of a viable matching procedure, which was then employed in linking the statistical sample.

The feasibility of matching is heavily dependent upon the size of the files involved. Theoretically, any record from an overlapping set in file "A" (A_i) and in file "B" (B_i) can be linked, but the number of matches or comparisons to be executed depends on the number of similar records in a subset to be searched; that is, "similarity" in terms of identifying information. If an individual's identifier is unique

tendances qui en découle ne peut toutefois mesurer que la **variation nette**, et non la **variation brute**. En outre, il est difficile de déduire les rapports de cause à effet à partir d'une analyse des tendances. Par contre, les données longitudinales facilitent l'étude de ces rapports, s'ils existent et qu'il sont liés dans le temps. Aussi les enquêtes réalisées auprès d'un échantillon constant ont-elles un avantage sur les "méthodes de mise à jour", car elles permettent de poser des questions spéciales en vue de préciser ou de déterminer la causalité.

Une compréhension de la causalité est essentielle à la formulation des politiques et au contrôle de leur incidence. Qu'une politique soit destinée à provoquer un changement ou à rester neutre, son objectif peut ne pas correspondre aux événements subséquents. On ne peut mesurer l'impact d'une politique sans tenir compte du temps et de la notion de la variation brute. À la limite, une variation nette nulle peut être considérée comme l'effet d'une politique neutre; en réalité, cependant, il s'est produit un effet positif sur une partie de la population et un effet négatif sur une autre; ces effets se soldent par des "transferts", qui sont fort différents d'une incidence nulle.

Si l'on songe à créer et à exploiter des données longitudinales, on doit d'abord prendre en compte certains facteurs d'ordre technique et pécuniaire. Évitions cependant de brûler les étapes: le couplage est une condition nécessaire de l'établissement d'un fichier longitudinal. Notre tâche immédiate consiste donc à déterminer la faisabilité et la rentabilité du couplage des dossiers, sans oublier pour autant les objectifs à long terme.

Appariement des dossiers de l'impôt et du recensement

Le coût et la faisabilité technique du couplage recensement-impôt ont été examinés dans une étude pilote qui portait sur un peu plus de 2,000 ménages de l'est de l'Ontario. L'étude a également fait progresser considérablement les techniques d'appariement utilisées dans le couplage de l'échantillon statistique.

Les possibilités d'appariement sont intimement liées au nombre de dossiers. Théoriquement, tous les dossiers de deux fichiers identiques "A" (A_i) et "B" (B_i) peuvent être appariés. Toutefois, le nombre de recherches et de comparaisons qui doivent être faites est fonction du nombre de dossiers identiques au sein du sous-groupe étudié, "identiques" s'entendant ici au sens de "renseignements d'identification". Seule l'utilisation d'un identificateur unique et sans erreur peut

and error-free, can the size of the input files be subordinated to other considerations. The census-taxation link, was not carried out under such ideal conditions. Consequently, size had to be controlled, and a sample had to be chosen.

The sample was selected from the Census (primary file), and the search file (secondary file) consisted of the personal identification file (RC-T) for the 1970 taxation year. This taxation year conforms to the income reference year for the 1971 Census, which is the 1970 calendar year. Identifying data are recorded only a few months apart. Most tax returns reflect the individual's status as it existed between January and May 1971, and Census data reflect the corresponding person's status as of June 1, 1971.

Identifying data, such as marital status, mailing address, and even name, change with time. Consequently, concordance of the chosen time frame is important, whenever such data have to be utilized for linkage applications.

The sample could have been selected from RC-T, and the Census could have been used as the search file, but a number of reasons dictated against such an approach. First of all, RC-T files permit the selection of individuals only. Thus, any linked file becomes a file of individuals; or conversely, a file of families can not be constructed with RC-T data as the primary source. Secondly, identifying information for Census records was not stored in machine-readable form. Because the complete secondary file has to be searched when attempting to match a sample, the utilization of the Census as the secondary file would have necessitated the capturing of identifying data for all Census records. When using the Census as the primary file and selecting a sample therefrom, additional data capture has to be carried out for the chosen sample only. With RC-T identifying information in machine-readable form, the data capture effort is minimized by our choice of primary and secondary files.

The Census sample was selected from the so-called 2B-file. This file is based on the long questionnaire, completed in 1971 by approximately one third of all households, and contains comprehensive socio-economic data, including income information. On the other hand, the 2A-file, which is based on the short Census questionnaire, is devoid of income data, and socio-demographic information is relatively scarce.

contribuer à éliminer le problème posé par le nombre des dossiers. Le couplage recensement - impôt ne s'est pas fait dans de telles conditions idéales. On a donc dû constituer un échantillon.

L'échantillon a été créé à partir des dossiers du recensement (fichier primaire); le fichier de recherche (fichier secondaire) était le fichier d'identification des particuliers (RC-I) pour l'année fiscale 1970. L'année fiscale est analogue à l'année de référence utilisée pour le revenu à l'occasion du recensement de 1971; elle équivaut à l'année civile 1970. Les données d'identification n'avaient été enregistrées qu'à quelques mois d'écart. En effet, la plupart des déclarations d'impôt portent sur le statut des particuliers entre janvier et mai 1971, alors que les données du recensement correspondent à leur statut au 1^{er} juin 1971.

Les données d'identification telles que l'état matrimonial, l'adresse postale et même le nom changent. Il est donc important que les cadres temporels choisis concordent si ces données doivent être utilisées dans l'appariement.

On aurait également pu constituer l'échantillon à partir des fichiers de RC-I et utiliser les résultats du recensement comme fichier de recherche. Cette approche n'a pas été retenue pour plusieurs raisons. Tout d'abord, les fichiers de RC-I ne permettent que le choix de particuliers. Il n'est donc pas possible de constituer un fichier de familles si l'on utilise les chiffres de RC-I comme source de base. Deuxièmement, les données d'identification des dossiers du recensement ne sont pas stockées sous une forme lisible par machine. Comme les recherches portent sur la totalité du fichier secondaire, il aurait fallu saisir les données d'identification de tous les dossiers du recensement. En revanche, en utilisant le recensement comme fichier primaire et en s'en servant pour constituer un échantillon, nous n'avons qu'à saisir les données d'identification des dossiers choisis. Enfin, comme les données d'identification du fichier de RC-I sont lisibles par machine, le choix des fichiers primaire et secondaire que nous avons fait limitait au minimum le travail de saisie des données.

L'échantillon du recensement a été constitué à partir du fichier 2B. Ce fichier est établi à partir de questionnaires détaillés, remplis en 1971 par environ un ménage sur trois; ces questionnaires contiennent d'abondantes données socio-économiques, et notamment des renseignements sur le revenu. Le fichier 2A, par contre est constitué à partir des questionnaires abrégés; il contient peu de renseignements caractéristiques socio-démographiques et aucune donnée sur le revenu.

The sample was selected as a clustered stratified sample. Enumeration areas (EA) were selected in the first stage. These EA's were stratified within provinces according to their metropolitan-urban-rural designation. Out of slightly over 42,000 EA's, 2,771 EA's comprised the first stage of the sample. These EA's were then subsampled by selecting every 12th household from a random start. This procedure yielded a sample of 16,357 individuals comprising about 33,000 households.

Once the households in the sample had been identified, all usable personal identifiers had to be captured from the Census questionnaire and made machine-readable. Unfortunately, questionnaires could not be located for 2,892 individuals, or 2.5% of the sample. This shortcoming removed 2,047 adults as potential matches.

The scope of identifying information was determined by the RC-T file content, for only data conceptually compatible and present on both files could be used for matching linkage decisions. Names, address, and month of birth had to be transcribed, and year of birth had to be verified. Sex and marital status, however, were taken from the machine-readable Census file and made numerically equivalent to codes in the RC-T file.

The data strings for matching consisted of name and address information, month and year of birth, sex, marital status, and where applicable, the first four characters of the given name of a person's spouse. Addresses on machine-readable RC-T records had to be reformatted and separated into such components as locality code, place name, postal code, box number, rural route number, civic number (house number) and street name.(8)

For subsequent file manipulations, a numeric Census identifier was carried to permit the merging of linked records to the census file **after all other identifiable information had been deleted.** Similarly, the RC-T account number and the Social Insurance number were carried on the RC-T data string to permit linkage to a separately maintained RC-T income file. These identifiers will also be needed when updating records over time without resorting to the use of conventional identifying information, because such information was deleted upon completion of the link between Census and RC-T records.

L'échantillon choisi était un échantillon stratifié par grappes. Dans une première étape, on a tiré les secteurs de dénombrement (SD). Ces SD ont été stratifiés par province en fonction de leur région d'appartenance (régions métropolitaines, urbaines et rurales). Il y avait au total 42,000 SD; de ce nombre, on en a choisi 2,771 dans une première étape. Ces SD ont ensuite été sous-échantillonnés, chaque 12e ménage étant retenu (le départ avait fixé au hasard). Nous avons ainsi obtenu un échantillon de 16,357 personnes formant environ 33,000 ménages.

Après que les ménages de l'échantillon aient été identifiés, tous les identificateurs personnels utilisables ont été saisis et rendus lisibles par machine. Malheureusement, nous n'avons pas pu trouver le questionnaire de 2,892 répondants, soit 2.5% de l'échantillon. 2,047 adultes ont ainsi été éliminés de l'appariement.

Le choix des renseignements d'identification était fonction du contenu des fichiers de RC-I; en effet, seules des données conceptuellement compatibles et présentes dans les deux fichiers pouvaient être utilisées dans le couplage. Le nom, l'adresse et le mois de naissance des répondants ont été transcrits; l'année de naissance a été vérifiée. Le sexe et l'état matrimonial, par contre, ont été directement tirés des fichiers lisibles par machine du recensement et rendus numériquement équivalents aux codes du fichier RC-I.

Les chaînes de données utilisées dans l'appariement comprenaient le nom et l'adresse, le mois et l'année de naissance, le sexe, l'état matrimonial et, le cas échéant, les quatre premières lettres du prénom du conjoint. Il a fallu changer la présentation des adresses figurant dans les dossiers de RC-I et décomposer celles-ci en plusieurs éléments: code de localité, nom de localité, code postal, numéro de case, numéro de route rurale, numéro de porte et nom de rue(8).

Pour faciliter les manipulations ultérieures, on a utilisé un identificateur numérique du recensement afin de permettre la fusion des dossiers couplés à ceux du fichier du recensement **après la suppression des autres renseignements identifiables.** De la même façon, le numéro de compte de RC-I et le numéro d'assurance sociale ont été intégrés aux chaînes de données de l'impôt afin que ces dossiers puissent être appariés à un fichier indépendant de RC-I. Ces identificateurs seront utilisés lors de la mise à jour des dossiers; en effet, les données d'identification habituelles ont été supprimées dès que le couplage des dossiers du recensement et de l'impôt a été terminé.

See footnote(s) at end of text.

Voir note(s) à la fin du texte.

Matching was carried out by computer, and a set of routines was programmed to decide whether or not a given comparison constituted a matched record pair. In a few cases, a number of record pairs was presented for manual assessment because alternative record pairs could not be declared positively to be matches or non-matches. A so-called hand-match had to be processed under these circumstances.

The primary aim of the decision-making process was speed combined with accuracy. To facilitate processing with these criteria in mind, the file was partitioned into 100 data blocks. The dividing lines were governed by alphabetic designators, which were made up of the first five characters of a person's surname.(9) Within these alphabetic blocks, files were sorted by surname, month of birth, and year of birth. A sort by locality code was also performed, but it was used only in those instances where names with high frequencies had to be restricted within geographical boundaries.

Routines to decide on the linkage status of record pairs were designed in the form of two separate rounds of interrogation. The first round attempted to locate all those matches which had a high probability of being true. The second round assessed the leftovers from the first round, but only those with a reasonable expectation of being in the tax universe. Second-round input thus contained fewer records than first-round input, since first-round matches were no longer present. Moreover, all records with a low probability of being a taxfiler were also omitted from second-round processing.

Out of approximately 33,000 households, 79,000 adults were eligible for matching, and 39,000 were matched during the first round. Out of the remaining 40,000, only 18,000 were re-entered and processed under round-two specifications. Each round, however, necessitated our entering of all RC-T records, which varied between 52,000 and 165,000 records depending on the subset to be linked.(10)

To further minimize the computer workload, each round was designed to make a positive disposition as to match or non-match as quickly as possible. Additional comparisons were made only if the affirmative or negative evidence was inconclusive. Thus, depending on the number of records in comparison space, not all steps in a given round were always followed when declaring a match. Along the way, several secondary records (candidates) were discarded if evidence indicated that their chance of becoming a match was extremely low.

See footnote(s) at end of text.

L'appariement des dossiers s'est fait par ordinateur. Des routines ont été spécialement conçues pour déterminer si les dossiers appariés étaient véritablement identiques. Dans quelques cas, les dossiers groupés ont été soumis à une vérification manuelle, l'ordinateur n'étant pas en mesure de déterminer hors de tout doute que les dossiers étaient identiques. L'appariement s'est alors fait manuellement.

L'objectif premier de ce processus de décision était d'accélérer l'appariement et d'en accroître la précision. Pour faciliter le traitement, le fichier a été divisé en 100 blocs de données; la division a été faite en fonction des cinq premières lettres du nom des répondants(9). Ces blocs alphabétiques étant constitués, les fichiers ont été triés par nom, mois de naissance et année de naissance. Un tri par code de localité a également été fait; on ne l'a toutefois utilisé que dans les cas où certains noms très fréquents devaient être restreints à un secteur géographique donné.

Les routines conçues pour déterminer le statut d'appariement des paires de dossiers comportaient deux séries d'interrogations distinctes. La première série avait pour but de trouver les appariements qui avaient une forte probabilité d'être bons. La deuxième série ne portait que sur les dossiers restants qui pouvaient raisonnablement se trouver dans l'univers de l'impôt. La deuxième série contenait moins de dossiers que la première, car tous les dossiers correctement appariés ne s'y trouvaient plus. De plus, les dossiers ayant une faible probabilité de correspondre à un contribuable étaient également omis.

Les 33,000 ménages de l'échantillon comprenaient 79,000 adultes pouvant faire l'objet d'un appariement; de ce nombre, 39,000 ont été appariés pendant la première série d'interrogations. Des 40,000 dossiers restants, seulement 18,000 ont été exploités pendant la deuxième série. À chaque occasion, toutefois, tous les dossiers de RC-I devaient être entrés, ce qui pouvait représenter de 52,000 à 165,000 dossiers, selon le sous-ensemble qui faisait l'objet du couplet(10).

Afin de réduire davantage le traitement informatique, chaque série d'interrogations a été conçue de façon à pouvoir déterminer le plus rapidement possible si l'appariement était bon. Les autres comparaisons n'intervenaient qu'en cas de doute. Ainsi, compte tenu du nombre de dossiers comparés, les diverses étapes d'une série n'avaient pas toutes lieu. Plusieurs dossiers (candidats) secondaires étaient éliminés s'ils avaient très peu de chances d'être appariés.

Voir note(s) à la fin du texte.

The likelihood of matching a record pair was based on empirical evidence from the pilot study.(11) Initially, a rough scoring scheme was introduced based on the frequency of certain variables but also considering their reporting reliability. A point score was accumulated depending on the agreement or disagreement of selected characteristics, such as sex, marital status, given names, locality code, place name and other address components, always provided surname, month of birth and year of birth agreed.(12) Secondary records with a very low score lost their eligibility at certain check points. If only one secondary record remained after elimination of unsuitable candidates, and if this record had attained a certain score, a match was declared. If the score of the single remaining record was low and not all steps in a round had been executed, additional comparisons between characteristics of the primary and secondary record were carried out. The final score thus attained determined the decision as to match or non-match.

Whenever more than one candidate remained for comparison space, the final decision as to which of these should be declared a match was usually decided on the basis of the highest score, provided the point spread was sufficient. Where the point spread was only marginal, the personal exemption was used as the final decision-making variable. If a tie could not be broken in this fashion, a visual examination or handmatch had to be carried out. There were only 46 hand matches in the entire project.

The explanations offered so far can only highlight the procedure and illustrate the principle involved. Some of these procedures will become meaningful later, when observations from linked record pairs will be discussed.

After subjecting the primary file to two rounds of matching routines, a set of matched record pairs and a set of non-matched primary records emerged. The secondary file was always used in its original size; i.e., records linked as part of a matched pair were not withdrawn. Consequently, a secondary record could enter a match several times. Obviously, only one of these duplicate matches could be true. Since it was relatively easy to locate duplicates on the basis of their unique RC-T identifiers, such conflicts were resolved after matching. A total of 87 duplicates had to be examined, and decisions with respect to their match status had to be made.

Les chances d'appariement d'une paire de dossiers étaient calculées sur des résultats empiriques tirés de l'étude pilote(11). Dès le départ, on s'est servi d'un mode de notation grossier fondé sur la fréquence de certaines variables et qui tenait également compte de la fiabilité des renseignements. Ainsi, on attribuait au dossier une note qui tenait compte du fait que certaines caractéristiques telles que le sexe, l'état matrimonial, les prénoms, le code de localité, le nom de localité et les autres éléments de l'adresse concordent ou non; le nom de famille, le mois de naissance et l'année de naissance devaient toujours concorder(12). Les dossiers secondaires ayant une très faible note étaient éliminés à certains points de contrôle. Si un seul dossier secondaire demeurait après l'élimination des autres candidats et si ce dossier avait une certaine note, on déclarait qu'il y avait appariement. Si la note du seul dossier restant était trop faible et que les comparaisons n'avaient pas toutes été faites, les caractéristiques du dossier primaire et du dossier secondaire étaient examinées plus à fond. C'est en fonction de la note finale ainsi obtenue qu'on déterminait s'il y avait appariement ou non.

Si plus d'un candidat demeuraient, la décision finale était généralement faite en fonction de la note la plus élevée, pourvu que l'écart entre les notes soit suffisant. Si l'écart n'était que minime, la décision finale était fondée sur les exemptions personnelles. En cas d'égalité, l'appariement était fait à la main. Le phénomène ne s'est présenté que 46 fois.

Les explications données jusqu'ici ne font ressortir que les points saillants et les principes de la méthode retenue. Certains aspects des procédures prendront leur pleine signification plus tard, quand nous examinerons les observations qui ont pu être faites à partir des dossiers couplés.

Après avoir soumis le fichier primaire aux deux routines d'appariement, nous avons obtenu un groupe de dossiers appariés et un groupe de dossiers primaires non appariés. Le fichier secondaire conservait toujours sa taille originale, car les dossiers appariés n'étaient pas retirés. Il s'en est donc suivi que certains dossiers secondaires ont été appariés plusieurs fois. De toute évidence, seul un de ces appariements était exact. Les dossiers de RC-I ayant un indicateur unique, il était relativement facile de trouver ceux qui avaient été retenus à plusieurs occasions; les 87 conflits qui se sont ainsi présentés ont été résolus après l'appariement.

It may have become apparent by now that not all decisions as to match status (match or non-match) were valid. Such a decision is always probabilistic. Borderline cases caused by large comparison sets with many similar records, or containing conflicting evidence due to faulty or missing data, may lead to false decisions. Thus, the decision to declare a certain match status can be **true or false.**

While one can accept the notion of a true match and a false match with relative ease, the notion of a true or false non-match is a bit more difficult to comprehend. Some non-matches are true; i.e., the primary record is of such a nature that it cannot be expected to have a counter-part in the tax file. Other non-matches are false; i.e., a match was not declared because a qualified candidate did not emerge, although the primary record belonged to a respondent who could be expected to have filed a tax return, given the institutional setting in 1970.

When a primary record remains unlinked, one cannot state explicitly what sort of comparison had been carried out, and what degree of agreement had been reached. However, the comparison and degree of variation in matched records will provide some indication of the data quality and its effect on the matching outcome. It is this data quality which is largely responsible for false non-matches.

On pourrait croire que les décisions quant au statut d'appariement (appariement ou non-appariement) n'étaient pas toutes justes. La chose est difficile à déterminer. En effet, comme les comparaisons portaient sur de grands ensembles de dossiers souvent similaires ou qui comportaient des renseignements contradictoires (certaines données étant inexactes ou absentes), il est raisonnable de croire que certaines décisions peuvent être erronées. Ainsi, les décisions relatives à un statut d'appariement donné peuvent être **vraies ou fausses.**

Bien qu'il soit relativement facile d'accepter qu'un appariement soit bon ou erroné, le concept de non-appariement bon ou erroné est un peu plus difficile à comprendre. Certains non-appariements sont justes: le dossier primaire est d'une telle nature qu'il ne peut pas avoir une contrepartie dans le fichier de l'impôt. Certains non-appariements sont faux: on a décidé qu'il n'y avait pas d'appariement parce qu'aucun candidat qualifié n'est ressorti, même si le dossier primaire appartenait à un répondant qui devait normalement avoir rempli une déclaration d'impôt, étant donné le cadre institutionnel de 1970.

Si un dossier primaire demeure non couplé, il est impossible de dire avec précision les comparaisons qui ont été faites et de définir le degré de concordance obtenu. Toutefois, la comparaison et le degré de variation des dossiers appariés nous renseignent sur la qualité des données et sur ses effets sur les résultats des appariements. C'est principalement à la qualité de ces données qu'on peut imputer les non-appariements erronés.

Methodological Review

The initial outcome of the matching exercise yielded 45,794 linked records. This number was later reduced because some of these matches contained duplicate secondary records, and only one of these matched pairs could be retained at best. Some matched pairs were judged to be false, and were converted to non-matches. Other matches, although considered to be true, could not be linked to RC-T income data. A total of 1,908 matched records was lost for the foregoing reasons, and 43,886 true matches were retained to form the base for a longitudinal income file.

The methodological review with reference to the efficacy of matching routines and agreement of variables is based primarily on the unedited linked file; i.e., it makes use of all matches regardless of duplicates or the truthfulness of matching decision.

The matching results can be summarized and cross-classified by two major characteristics, namely the round in which the match occurred, and the type of decision. This type will be classified as "unique" or "multiple", where multiple implies that more than one candidate had to be considered for the final matching decision. Unique decisions, on the other hand, were based on the point accumulation of one secondary record vis-à-vis one primary record. Approximately half of all decisions involving unique matches were made after unsuitable candidates had been eliminated. The results are summarized in Text Table I.

Méthodologie

L'appariement a produit 45,794 dossiers couplés. Ce nombre a par la suite diminué quelque peu, certains appariements comportant le même dossier secondaire. Par ailleurs, on a estimé que certains appariements étaient erronés et on a décidé qu'il y avait non-appariement. D'autres appariements, enfin, n'ont pas pu être couplés aux données sur le revenu de RC-I en dépit du fait qu'ils aient été bons. Au total, 1,908 dossiers appariés ont été éliminés. Le fichier longitudinal sur le revenu s'appuie donc sur 43,886 appariements véritables.

L'étude méthodologique de l'efficacité des routines d'appariement et de la qualité de la concordance des variables s'appuie principalement sur le fichier couplé non contrôlé; on se sert donc de tous les appariements, qu'ils soient justes ou faux et même si certains dossiers sont repris plusieurs fois.

Les résultats de l'appariement peuvent être présentés de façon sommaire et classés en fonction de deux principales caractéristiques: la série de comparaisons pendant laquelle l'appariement s'est fait et le genre de décision. Cette variable est elle-même divisée en deux catégories: le groupe "unique" et le groupe "multiple", ce dernier correspondant aux cas où plusieurs candidats ont dû être pris en compte. Les décisions à caractère unique, pour leur part, correspondent aux situations où il n'y avait qu'un seul dossier secondaire par dossier primaire. Environ la moitié des décisions relatives aux appariements à caractère unique ont été faites après l'élimination des candidats indésirables. Les résultats sont présentés de façon sommaire au tableau explicatif I.

TEXT TABLE I. Matches by Time of Occurrence of Decision Type in Census-RC-T Match, 1971

TABLEAU EXPLICATIF I. Appariements selon le moment du type de décision, appariement recensement-RC-I, 1971

	First round	Second round	Total
	Première série	Deuxième série	
Unique decision - Décision unique	37,940	4,852	42,792
Multiple decision - Décision multiple	1,168	1,834	3,002
Total	39,108	6,686	45,794

It can be seen that the majority of matched pairs was created in the first round, and that most of these first-round matches were based on unique decisions. It should be recalled that the design aimed at bringing about speedy and accurate decisions. With first-round matching conditions being more stringent than second-round ones, and with unique decisions preceding multiple decisions; i.e., utilizing less computer time, one can substitute "accuracy levels" for "round" and "computer time levels" for "decisions types". When expressing each cell as a percentage of the grand total, the following results are obtained, as can be seen from Text Table II.

TEXT TABLE II. Accuracy Levels and Computer Time Levels for the Census-RC-T Match, 1971

TABLEAU EXPLICATIF II. Niveau de précision et temps d'ordinateur, appariement recensement-RC-I, 1971

	Great accuracy	Moderate accuracy	Total
	Grande précision	Précision moyenne	
	per cent - pourcentage		
Fast solution - Solution rapide	82.8	10.6	93.
Slow solution - Solution lente	2.6	4.0	6.
Total	85.4	14.6	100.

Sufficient speed and accuracy can be attributed to 82.8% of all matches, and 2.6% were attained with great accuracy but involving relatively more computer time. Thus, great accuracy can be claimed for 85.4% of all matches. Moderate accuracy can be attached to 14.6% of all matched pairs, with 10.6% of this group being classified as "fast". This characteristic is a relative attribute within the second round only. Remember that all second-round matches had passed through the first round. They have used more computer time than the 2.6% of all matches classified as slow but having great accuracy.

One of the more complex and time-consuming tasks centered on the utilization of mailing addresses. RC-T addresses were in machine-readable form, but the record format was such that extensive reformatting and decomposition of mailing addresses was required. Items such as rural route numbers, box numbers, house numbers (civic numbers), and apartment numbers(13) had to be iso-

See footnote(s) at end of text.

À la lecture du tableau, on peut voir que la majorité des appariements se sont faits pendant la première série de comparaisons et qu'ils reposaient pour la plupart sur des décisions à caractère unique. Il convient de rappeler que le programme avait été conçu de façon à produire rapidement des décisions précises. Les conditions d'appariement de la première série de comparaisons étant plus sévères que celles de la seconde et les décisions uniques précédant les décisions multiples (en ce sens qu'elles consomment moins de temps d'ordinateur), on peut remplacer les expressions "série" par "niveau de précision" et "genre de décision" par "temps d'ordinateur". Si l'on exprime chaque case par un pourcentage du total, on obtient les résultats du tableau explicatif II.

On peut dire que 82.8% des appariements répondent à des critères de rapidité et de précision suffisants et que 2.6% des appariements ont atteint une grande précision, l'opération ayant toutefois nécessité un peu plus de temps d'ordinateur. Ainsi, 85.4% des appariements ont une grande précision. D'autre part, 14.6% des appariements ont une précision moyenne, l'opération s'étant faite rapidement dans 10.6% des cas. Cette caractéristique ne constitue toutefois qu'une qualité relative dans la deuxième série de comparaisons. On se rappellera en effet que les appariements de la deuxième série ont tous passé par la première. Ils ont donc utilisé plus de temps d'ordinateur que les appariements "lents" mais très précis (2.6%).

L'une des tâches les plus complexes et les plus longues avait trait à l'utilisation des adresses postales. Les adresses de RC-I étaient lisibles par machine, mais présentées de telle façon qu'on a dû les décomposer et les reformater. Des éléments tels que le numéro de route rurale, le numéro de case, le numéro de porte (numéro de voirie) et le numéro d'appartement(13) ont dû être isolés. Par ailleurs, les noms

Voir note(s) à la fin du texte.

ated. Apart from numeric address information, place names and street names had to be identified, and often a distinction between street name and building name was needed. Streets, of course, can also be identified by numerics. Moreover, street designations such as "avenue", "lane", "drive" and about 5 others occur in either one or the other of our official languages, and concordance between Census and RC-T with respect to language use for a given record is lacking.

The use of address was further complicated by conceptual differences and time-frame variations. With Census day being June 30th, and tax returns having been filed mostly by April 30th, any change of address which took place between the time of filing one's tax return and completing one's Census questionnaire resulted in discrepant address information. The conceptual difference, on the other hand, arises from RC-T requests for **mailing address** and census requests for **place of residence**. While these two definitions agree in most instances, they differ in those cases where individuals decide to have their mail directed to a place other than their residence. Definition of "address" is particularly troublesome in rural areas where post-office boxes and rural routes prevail, and where Census "addresses" consist of lot and concession number and other designators pertaining to the subdivision of land. Nevertheless, attempts were made to utilize addresses as best as could be ascertained.

All RC-T records had mailing addresses available, although some of these records posed problems when attempting to decompose and reformat their address fields. Out of 5,794 matched records, only 38,093 (83.2%) had a usable Census mailing address. However, the use of address as part of the matching routines was restricted to records where other information failed to identify a match conclusively. Consequently, the address was used to a limited extent, namely in 3,793 instances.

The potential usefulness of address information is of methodological interest. Thus, address components were investigated with respect to availability and agreement. A component was available if it occurred on both segments of a matched record. Agreement was judged as partial, if such partial agreement was accepted in matching routines. Conversely, disagreement implies that a certain characteristic did not contribute anything to the point score leading to a matching decision.

Agreement or disagreement may be consistent or inconsistent. Agreement is consistent with true matches, but inconsistent

localité et de rue ont dû être identifiées, et l'on a souvent dû faire une distinction entre le nom de la rue et celui de l'édifice. Les rues, en outre, peuvent également être désignées par un numéro. Enfin, on a relevé une bonne quinzaine d'appellations telles que "avenue", "chemin" et "promenade" dans les deux langues officielles, sans qu'il y ait concordance entre les dossiers du recensement et ceux de RC-I.

L'utilisation de l'adresse se compliquait en outre du fait qu'il y avait entre les deux fichiers des différences d'ordre conceptuel et temporel. Par exemple, comme la journée du recensement était le 1^{er} juin et que les déclarations d'impôt avaient été remplies pour la plupart avant le 30 avril, les changements d'adresse survenus entre ces deux dates ont été à l'origine de divergences. Du point de vue conceptuel, d'autre part, on observe que les dossiers de RC-I sont fondés sur l'adresse postale, alors qu'au recensement, on demande le lieu de résidence. Bien que ces deux définitions coïncident dans la plupart des cas, elles diffèrent quand certaines personnes dirigent leur courrier ailleurs qu'à leur résidence. La définition de l'adresse est particulièrement compliquée dans les régions rurales où les casiers postaux et les routes rurales dominent, et où l'adresse aux fins du recensement consiste en un numéro de lot ou de concession et en d'autres renseignements relatifs à la subdivision du sol. On a néanmoins tenté d'utiliser au mieux les adresses disponibles.

Tous les dossiers de RC-I avaient une adresse postale; certains dossiers ont néanmoins posé des problèmes quand il a fallu décomposer et reformuler leur adresse. Des 45,794 dossiers appariés, seulement 38,093 (83.2%) avaient une adresse postale du recensement utilisable. Toutefois, l'utilisation de l'adresse dans les routines d'appariements se limitait aux dossiers dans lesquels les autres renseignements n'avaient pas donné un appariement sûr. L'adresse a donc été utilisée dans un nombre limité de cas (3,793).

Les possibilités offertes par l'adresse ont plutôt un intérêt méthodologique. C'est pour cette raison qu'on a étudié la disponibilité et la concordance des éléments de l'adresse. Un élément était qualifié de "disponible" s'il paraissait dans les deux segments d'un dossier apparié. La concordance était dite partielle si elle avait été acceptée par les routines d'appariement. À l'inverse, on a établi qu'il y avait divergence si une caractéristique donnée n'augmentait pas la note attribuée à un dossier pendant le processus d'appariement.

La convergence ou la divergence peuvent être cohérentes ou incohérentes. Ainsi, la convergence est cohérente dans les appariements justes, mais

with false matches, where it is a random event. Disagreement should be expected in false matches; thus, it is consistent, whereas disagreement in true matches indicates the unreliability of such a characteristic. On the other hand, it shows that disagreement in isolation may not prevent a valid match from being declared.

Address components are compiled in Text Table III together with their consistent and inconsistent frequency of occurrence. Finelocality codes are included therein since they are quasi addresses. They are not independent of place name. Finelocality codes at the three-digit level (FINELOC3) embrace larger areas than those at the five-digit level (FINELOC5). Usually, one associates five-digit codes with **municipality** and three-digit codes with **county** or **Census division**.

The relatively large area covered by locality codes and place name, increases the chance of agreement for false matches as is revealed by 627 inconsistent agreements of three-digit locality codes and 380 chance agreements for place name.

The lack of agreement of address components for true matches is possible because actual use of address components was made for only 10% of all matches. It shows, however, that greater reliance on addresses, as they are now supplied, would affect match results adversely.

Two possible approaches can be outlined to circumvent the detrimental effect of inconsistent address components. First, one can try to increase the data quality of addresses, if their use is unavoidable. Secondly, one can obtain other types of data which contribute appreciably to a matching decision, provided such data can be obtained easily, and can be expected to have a higher degree of reliability than address components.

To put the foregoing statement into better perspective, the consistency status of other matching characteristics will be examined next. These characteristics are of a personal nature and they include month of birth (MOB), year of birth (YOB), marital status, which is expressed in coded form and contains **single**, **married**, **separated**, **divorced**, and **widowed**. Sex is another charac-

incohérente dans les appariements erronés, elle est dictée par le hasard. Les cas de divergence se présentent dans les appariements erronés; il y a alors divergence cohérente. revanche, les cas de divergence observés dans les appariements justes font ressortir le manque de fiabilité de cette caractéristique. Prise isolément, la divergence n'empêchait pas toutefois qu'un appariement valide puisse se faire.

Les éléments de l'adresse ainsi que la fréquence de cohérence et d'incohérence sont présentés au tableau explicatif III. Les codes de localité y sont présentés, car ils correspondent pratiquement à une adresse. Ils ne sont pas en rapport avec le nom de la localité. Les codes de localité à trois chiffres couvrent des régions plus grandes que les codes à cinq chiffres. De façon générale, les codes à cinq chiffres correspondent à la **municipalité**, et les codes à trois chiffres, au **comté** ou à la **division de recensement**.

La superficie couverte par les codes et les noms de localité étant relativement grande, elle accroît les risques de convergence des dossiers erronés; il y a en effet eu 627 cas de convergence incohérentes pour les codes de localité à trois chiffres et 380, pour les noms de localité.

L'absence de convergence entre les éléments de l'adresse des appariements justes vient de ce que les éléments de l'adresse n'ont été véritablement utilisés que dans 10% des appariements. Le phénomène montre néanmoins que, si l'on s'appuyait davantage sur les adresses telles qu'elles sont présentées à l'heure actuelle, la qualité de l'appariement en souffrirait.

Pour contourner les effets négatifs de l'incohérence des éléments de l'adresse, on peut faire appel à deux approches. D'une part, on peut tenter d'améliorer la qualité des données des adresses si l'on est obligé d'y avoir recours. D'autre part, on peut obtenir d'autres données qui jouent un rôle important dans l'appariement, pourvu que ces données soient facilement accessibles et qu'elles aient un niveau de fiabilité supérieur à celui des éléments de l'adresse.

Pour bien situer les choses, nous examinons maintenant le statut de cohérence d'autres caractéristiques d'appariement. Ces caractéristiques sont: le mois de naissance (MDN), l'année de naissance (ADN), l'état matrimonial (**célibataire**, **marié**, **séparé**, **divorcé** et **veuf**), le sexe, la caractéristique appelée "prénom et initiales" et, enfin, le "CONJOINT". Cette appellation sert à désigner "les quatre premiers caractères de

TEXT TABLE III. Address Components by Consistency Status and Agreement Type for the Census-RC-T Match, 1971

TABLEAU EXPLICATIF III. Éléments de l'adresse selon le statut de cohérence et le type de convergence, appariement recensement-RC-T, 1971

Agreement type(1) - Type de convergence(1)						
Data item - Élément	Inconsistent usage				Consistent usage	
	Usage incohérent				Usage cohérent	
	A		B		C	
	No. - nbre	%	No. - nbre	%	No. - nbre	
Five-digit locality code - Code de localité à cinq chiffres	116	0.7	3,972	25.4	11,286	72.9
Three-digit locality code - Code de localité à trois chiffres	627	1.4	3,080	6.7	40,925	89.4
Place name - Nom de localité	380	1.0	6,793	17.8	29,852	78.5
Street name (first eight characters) - Nom de rue (huit premiers caractères)	55	0.2	5,690	17.8	25,062	78.4
Civic number - Numéro de voirie	43	0.2	3,252	10.7	26,068	86.1
Box number - Numéro de case	2	1.3	18	11.7	130	84.4
Rural route - Route rurale	6	1.1	17	3.3	490	94.8
Agreement type(1) - Type de convergence(1)						
Consistent usage				Intermediate usage		Number of items available for comparison(2) Nombre d'éléments disponibles pour fins de comparaison(2)
Usage cohérent				Usage intermédiaire		
D		E		F		
	No. - nbre	%	No. - nbre	%	No. - nbre	%
Five-digit locality code - Code de localité à cinq chiffres	297	1.9	-	-	-	15,671
Three-digit locality code - Code de localité à trois chiffres	1,162	2.5	-	-	-	45,794
Place name - Nom de localité	1,025	2.7	-	-	-	38,050
Street name (first eight characters) - Nom de rue (huit premiers caractères)	1,083	3.4	48	0.1	11	0.1
Civic number - Numéro de voirie	936	3.1	-	-	-	-
Box number - Numéro de case	4	2.6	-	-	-	-
Rural route - Route rurale	4	0.8	-	-	-	512

(1) Agreement types A to F are defined as follows:

- A. Characteristic agrees although match is false.
- B. Characteristic disagrees although match is true.
- C. Characteristic agrees and match is true.
- D. Characteristic disagrees and match is false.
- E. Characteristic (variable) agrees partly for true match.
- F. Characteristic agrees partly for false match.

(1) Les types de convergence A à F sont définis comme suit:

- A. Caractéristique convergente, appariement erroné.
- B. Caractéristique divergente, appariement juste.
- C. Caractéristique convergente, appariement juste.
- D. Caractéristique divergente, appariement erroné.
- E. Caractéristique (variable) partiellement convergente, appariement juste.
- F. Caractéristique partiellement convergente, appariement faux.

(2) All percentages use "number of available items" as a base.

(2) Tous les pourcentages sont fondés sur le "nombre d'éléments disponibles".

teristic to be examined for matching consistency. The discussion also includes a characteristic called "First Name and Initials", and finally "SPOUSE". This last designation stands for "the first four characters of a spouse's given name". RC-T uses "commonly used given name", but we had to infer from the Census record which given name to choose, if more than one was stated in full.

As can be seen from Text Table IV, sex has the greatest incidence of inconsistent usage for false matches. This finding is not surprising since the chance of random agreement for "sex" is quite high. Similarly, marital status ranks second for being inconsistent in false matches. It is even more pronounced in terms of disagreement for true matches. This fact indicates that marital status is neither reliably reported nor does it have a great discriminating power. Sex, on the other hand, also lacks discriminating power, but is very reliably reported since only 80 records disagree on sex, yet are true matches.

Month of birth (MOB) occupies a middle ground in discriminating power, as indicated by 1,144 (2.5%) inconsistent agreements for false matches. Its reporting reliability can be judged on the basis of 1,604 disagreements, although the match is true.

Year of birth (YOB) has greater discriminating power than MOB, which is obvious on *a-priori* grounds since the probability of agreeing by chance is 1:12 for MOB whereas it is only 1:50 for YOB over the expected range, although "bunching" in prime age groups makes for lack of uniformity.

To enhance the discriminating power of date-of-birth information, day, month, and year in combination would improve the results appreciably. Reporting errors, as reflected in the B-groups of Text Table IV should remain close to those experienced for MOB, although special care during collection and processing may further improve date-of-birth data.

First names and initials have excellent discriminating power, as can be inferred from a failure rate of 1.2%, or 532 inconsistent records in group A. The quality of this data item, however, is affected by inconsistencies that go beyond simple "errors". Depending on the type of document, formal first names are replaced with short forms or adopted unofficial first names.

prénom du conjoint". RC-I utilise le "prénom usuel"; pour les dossiers du recensement, toute fois, nous avons parfois dû choisir le prénom s'il n'en avait donné plus d'un.

Comme on peut le voir à la lecture du tableau explicatif IV, c'est le sexe qui revient le plus souvent dans les appariements erronés. La chose n'est pas surprenante si l'on songe que les chances de convergence fortuite de la caractéristique "sexe" sont assez élevées. L'état matrimonial vient à cet égard au deuxième rang. Sa présence est d'ailleurs encore plus prononcée dans les cas de divergence d'appariement justes. Cela montre bien qu'on ne peut pas s'appuyer sur la qualité de la déclaration de l'état matrimonial et que cette caractéristique n'a pas un pouvoir discriminant élevé. Le sexe, pour sa part, a lui aussi un faible pouvoir discriminant; en revanche, il est très bien rapporté, puisqu'il n'y a divergence que dans 80 dossiers (les appariements étant toutefois justes).

Le mois de naissance (MDN) occupe une position intermédiaire du point de vue du pouvoir discriminant; on observe en effet qu'il y a eu 1,144 cas (2.5%) de convergence incohérente (appariements erronés). La fiabilité du MDN peut être appréciée au fait qu'il y a eu 1,604 divergence dans des cas d'appariements justes.

L'année de naissance (ADN) a un meilleur pouvoir discriminant que le MDN. La chose est évidente puisque la probabilité de convergence fortuite est de 1:12 dans le cas du MDN et de seulement 1:50 dans celui de l'ADN pour les groupes d'âge étudiés, le grand nombre de répondants d'âge moyen ayant néanmoins un effet adverse à cet égard.

Pour accroître le pouvoir discriminant de la date de naissance, il faudrait utiliser le jour, le mois et l'année de naissance. Les erreurs de déclaration mises en évidence par les groupes du tableau explicatif IV demeureraient voisines de celles qui s'observent avec le MDN; toutefois, cela arriverait peut-être à améliorer les données sur la date de naissance en portant une attention spéciale au processus de collecte et de traitement.

Les prénoms et les initiales ont un excellent pouvoir discriminant; en effet, le taux d'échec n'a été que de 1.2% (532 dossiers incohérents dans le sein du groupe A). La qualité de ces données est toutefois, liée aux incohérences qui vont au-delà de la simple "erreur". Compte tenu de la nature du document, les prénoms officiels sont remplacés par des prénoms usuels. Souvent, le prénom intermédiaire s'est substitué au prénom

TEXT TABLE IV. Personal Characteristics and Variables by Consistency Status and Agreement Type for the Census-RC-T Match, 1971

TABLEAU EXPLICATIF IV. Caractéristiques et variables personnelles selon le statut de cohérence et le type de convergence, appariement recensement-RC-I, 1971

Agreement type(1) - Type de convergence(1)						
Item - Élément	Inconsistent usage			Consistent usage		
	Usage incohérent			Usage cohérent		
	A		B		C	
	No. - nbre	%	No. - nbre	%	No. - nbre	%
Month of birth - Mois de naissance	1,144	2.5	1,604	3.5	42,401	92.6
Year of birth - Année de naissance	821	1.8	-	-	41,619	90.4
Marital status - État matrimonial	1,608	3.5	4,800	10.5	39,205	85.6
Sex - Sexe	1,766	3.8	80	0.2	43,925	95.8
First names initials - Prénoms initiales	535	1.2	2,102	4.6	28,601	62.5
First four characters of spouse's first name - Quatre premiers caractères du prénom du conjoint	290	1.0	1,765	5.9	25,890	87.2
Agreement type(1) - Type de convergence(1)						
	D		E		F	
	No. - nbre	%	No. - nbre	%	No. - nbre	%
Month of birth - Mois de naissance	645	1.4	-	-	-	45,794
Year of birth - Année de naissance	-	-	2,386	5.2	968	2.1
Marital status - État matrimonial	181	0.4	-	-	-	45,794
Sex - Sexe	23	0.1	-	-	-	45,794
First names initials - Prénoms initiales	342	0.7	13,302	29.0	910	2.0
First four characters of spouse's first name - Quatre premiers caractères du prénom du conjoint	238	0.5	-	-	-	29,874

1) See footnote (1) Text Table III.

1) Voir note (1) du tableau explicatif III.

2) All percentages use "number of available items" as a base.

2) Tous les pourcentages sont fondés sur le "nombre d'éléments disponibles".

Often, the middle name has become the commonly used given name, but lack of reporting consistency between data sources makes it difficult to use this data item to its fullest capacity.

Comments with reference to first names also apply to SPOUSE, since it is a first-name-derived data item. Consequently, it is afflicted with similar strengths and weaknesses.

Other inferences could be made from Text Table IV, but these are left to the reader. The results of the methodological evaluation will now be summarized before proceeding with the analysis of matching results.

Speed and accuracy of the matching procedure are governed by the number of variables used and by the discriminating power attributed to these variables. Discriminating power is mainly a property of the "uniqueness" of the variable, but the quality of such a variable or characteristic in terms of reporting and processing reliability is crucial.

While the choice of data is dependent on circumstances often beyond the control of the statistical agency, especially if administrative data are used, the quality of the data can often be ameliorated by special processing. This procedure is particularly applicable to the primary file, which usually consists of a sample.

As data linkage becomes a more widely-used process than heretofore employed, choice of data and data quality should be improved by **appropriate collection procedures**.

Computer-programmed decisions to generate linked records are highly dependent on the amount of data to be compared (scale). As files increase in size, comparisons increase exponentially. Thus, large-scale matching operations may become prohibitively costly or highly inaccurate. Advances in computer technology may help to increase technical feasibility and thus have to be evaluated periodically. Computer processing, of course, is also affected by the programmed routines, which in turn have to make allowances for the type of data available. The collection of more suitable data for record linkage applications thus promises to yield the greatest benefits in terms of the overall effectiveness of linkage applications.

usuel; toutefois, le manque d'uniformité des méthodes de déclaration nous empêche de tirer pleinement profit de ces données.

La situation étant analogue dans le cas concept du CONJOINT, cette caractéristique a des avantages et des inconvénients analogues.

D'autres conclusions pourraient être tirées du tableau explicatif IV; nous laisserons ce soin au lecteur. Avant d'analyser les résultats de l'appariement, nous présenterons sommairement les résultats de l'évaluation de la méthodologie de l'opération.

La vitesse et la précision de l'appariement sont liées au nombre de variables utilisées et au pouvoir discriminant attribué à chacune d'elles. Le pouvoir discriminant tient surtout à l'unicité de la variable, alors que la qualité de cette dernière est fonction de la fiabilité de la déclaration et du traitement des données.

Bien que le choix des données dépende des circonstances sur lesquelles l'organisme statistique n'a souvent aucun contrôle - surtout s'il s'agit de données administratives - on peut fréquemment améliorer la qualité des renseignements au moyen d'un traitement spécial. Cela est particulièrement vrai dans le cas du fichier primaire, qui se présente le plus souvent sous la forme d'un échantillon.

Au fur et à mesure que le couplage des données se répandra, l'utilisation de **procédures de collecte appropriées** devrait améliorer le choix et la qualité des données.

Le couplage par ordinateur de dossiers intimement liés au volume des données qui doivent être comparées. Au fur et à mesure que la taille des fichiers s'accroît, le nombre des comparaisons augmente de façon exponentielle. Ainsi, les opérations d'appariement à grande échelle peuvent occasionner des coûts prohibitifs ou devenir hautement imprécises. Les progrès réalisés dans le domaine de l'informatique pourront néanmoins contribuer à en accroître la faisabilité; la situation devrait donc être évaluée périodiquement. Le traitement informatique dépend évidemment des routines programmées; à leur tour, celles-ci doivent tenir compte de la nature des données disponibles. C'est donc la collecte de données mieux adaptées au couplage des données qui est susceptible d'offrir les résultats les plus prometteurs en cette matière.

Income Reporting on Matches and Non-matches

All matched records had to be combined in RC-T income data strings, and identifying information, except social insurance number and account number (REDID), was removed at this stage. As was mentioned before, some matched records could not be merged with their income portions of the RC-T file due to file updating problems. Therefore, the analysis which follows is based on a slightly reduced universe.

The sample as selected contained 116,380 records and 79,181 of these were adults. They are defined in the Census as being 15 years of age or older. Since income questions had been asked of adults only, children were excluded from matching considerations. This restriction does not preclude their being retained in households or families for analytical purposes with respect to family size or composition.

Attempts to link these 79,181 adults to RC-T files yielded 45,665 matched pairs for which income information was available, but 43,886 were accepted as true matches. This decision was made without consideration of income reporting.

The group of 1,779 false matches was returned to the non-matches segment of the file; however, false matches remain identifiable by way of a code.

The non-match set is made up of 33,516 original non-matches, i.e., those which were rejected from computer decisions, and 1,779 "inverted" non-matches, namely former matches classified as "false".

In terms of **income sources**, as reported in the Census, matches and non-matches differ markedly. Although 1,173 matched records showed no Census income, most non-matched records are associated with non-matches, namely 19,938.

Matched records show a high incidence of earned income with 88.9% of all income recipients having this type as its major source. The remainder, 11.1% made up of 12,888 records, have non-earned income as a major source.

Non-matches, on the other hand, have only 11.1% of income recipients as earned-income recipients by source categories, whereas 46.7% of non-matched income recipients report non-earned income types as their major source. Further detail can be obtained from Table 1.

Déclaration du revenu - Dossiers appariés et non appariés

À cette étape du travail, on a groupé les dossiers appariés aux chaînes de données sur le revenu de RC-I, et les données d'identification (exception faite du numéro d'assurance sociale et du numéro de compte) ont été éliminées. Comme on l'a déjà vu, des problèmes de mise à jour du fichier nous ont empêchés de fusionner certains dossiers appariés aux données sur le revenu de RC-I. L'analyse qui suit porte donc sur un univers légèrement diminué.

L'échantillon choisi comportait 116,380 dossiers; 79,181 d'entre eux correspondaient à des adultes (personnes qui, au sens du recensement, ont 15 ans et plus). Comme les questions sur le revenu n'étaient posées qu'aux adultes, les enfants ont été exclus de l'appariement. On les a néanmoins conservés dans les ménages ou les familles à des fins analytiques (taille ou composition de la famille).

Les tentatives de couplage de ces 79,181 adultes aux fichiers de RC-I ont donné 45,665 appariements pour lesquels des renseignements sur le revenu étaient disponibles; seulement 43,886 d'entre eux étaient bons. Cette décision a été prise sans tenir compte du revenu.

Les 1,779 appariements erronés ont été renvoyés au segment des non-appariements du fichier. On peut néanmoins les reconnaître au code qui leur a été attribué.

Le groupe des non-appariements se compose des 33,516 dossiers non appariés par l'ordinateur et des 1,779 appariements rejetés.

Les appariements et les non-appariements diffèrent sensiblement du point de vue des **sources de revenu** déclarés au recensement. Bien que 1,173 dossiers appariés ne contenaient aucune donnée sur le revenu, la plupart des dossiers à revenu nul (19,938) étaient associés à des non-appariements.

Dans la majorité des dossiers appariés (88.9%), les répondants tiraient leur principale source de revenu d'un revenu gagné. Dans les 4,738 dossiers restants (11.1%), la principale source de revenu était un revenu non gagné.

En ce qui concerne les non-appariements, d'autre part, seulement 53.3% des bénéficiaires d'un revenu avaient comme principale source de revenu un revenu gagné; les 46.7% restants ne tiraient pas leur principale source de revenu d'un revenu gagné. On trouvera d'autres renseignements à ce sujet dans le tableau 1.

It will be recalled that many non-matches do not constitute a "failure". Most non-matches represent a correct decision because the Census record, which reflects such an outcome, belongs to a person who could not have been expected to file a tax return due to the absence of income for taxing purposes. Consequently, a link to a tax record is impossible under these circumstances.

The success of the matching project can best be judged in terms of "all true matches" out of the "estimated number of taxfilers" expected to coincide with the census sample. The estimated number of taxfilers consists of all "true matches" and all "false non-matches", and amounts to 47,970 records. This result can be expressed as a match rate of 43,886/47,970 or 91.5% of the overlapping universe. The underlying data for this match rate are presented in Table 15.

The taxfiler universe in Table 15 has been estimated on the basis of Census information, including pertinent family and dependency relationships. Alternatively, a taxfiler rate can be calculated from published Revenue Canada information. Applying this rate to all adults in our sample, an estimated number of taxfilers emerges, and matching success can be judged against this subset. Match Rate II in Table 20 shows generally a greater degree of success. While it confirms the approach used in Table 15, it should not be used as the ultimate criterion for judgment.

The failure rate in Table 20 indicates the percentage of false non-matches out of all adults in the sample. Its complement, the success rate, includes all valid decisions; in other words, true non-matches as well as true matches are successful outcomes. Success rates vary between 92.7% and 95.6% for provinces, with the weighted Canadian rate at 94.8%.

The match rate, however, is more relevant than the success rate in view of the intended use of the data. A match rate of 91.5% implies a non-match rate of 8.5%. An attempt will now be made to assess the shortfall in income due to non-matches. There are 4,084 non-matches which should have been matched (false non-matches). They account for \$25.034 million in the sample which represents 9.1% of total income.

There are also slightly over 11,000 non-matches with small amounts of income, but individuals presented by these records may

On se rappellera que bon nombre de non-appariements ne constituent pas un "échec". La plupart des non-appariements correspondent à une bonne décision, le dossier du recensement correspondant à une personne qui n'a pas rempli déclaration d'impôt faute de revenus aux fins de l'impôt. Ces dossiers ne peuvent manifester pas être couplés à ceux de l'impôt.

La meilleure façon de mesurer le succès du projet d'appariement consiste à comparer "nombre total d'appariements justes" au "nombre estimatif de contribuables" qui devrait coïncider avec l'échantillon du recensement. Le nombre estimatif de contribuables comprend les "appariements justes" et les "non-appariements erronés" et s'élève à 47,970 dossiers. Ces résultats donnent donc un taux d'appariement de 43,886/47,970 soit 91.5%. Les données utilisées pour faire le calcul sont présentées au tableau 15.

L'univers des contribuables du tableau 15 a été estimé à partir de chiffres du recensement et notamment de renseignements sur les familles et les personnes à charge. Le pourcentage de contribuables peut également être calculé à partir de renseignements publiés par le Canada. En appliquant ce taux à l'ensemble des adultes de notre échantillon, on obtient un nombre estimatif de contribuables en fonction duquel on peut évaluer le succès de l'appariement. Le taux d'appariement II présenté au tableau 20 correspond à un meilleur degré de réussite. Bien que l'approche utilisée dans le tableau 15 se trouve confirmée, ce taux ne devrait pas servir de critère de jugement ultime.

Le taux d'échec du tableau 20 représente la proportion des non-appariements erronés en rapport à l'ensemble des adultes de l'échantillon. Son complément, le taux de réussite, sert à désigner l'ensemble des décisions valides. D'autres termes, les non-appariements justes représentent une réussite au même titre que les appariements justes. Les taux de réussite oscillent entre 92.7% et 95.6% d'une province à l'autre, le taux canadien pondéré s'établissant à 94.8%.

Si l'on tient compte de la destination des données, le taux d'appariement est néanmoins plus pertinent que le taux de réussite. Un taux d'appariement de 91.5% suppose un taux de non-appariement de 8.5%. Nous essaierons maintenant d'évaluer la différence en moins au niveau des revenus attribuable aux non-appariements. Il y a eu 4,084 dossiers non appariés, mais qui auraient dû l'être (non-appariements erronés). Ces dossiers représentent dans l'échantillon \$25.034 millions, soit 9.1% du revenu total.

On compte également un peu plus de 11,000 dossiers non appariés correspondant à de faibles revenus pour lesquels il n'était peut-être

have been required to file a tax return in 1970. Thus they are classified as "true non-matches". They generate \$11.129 million or 4.0% of total income in the sample. The combined shortfall in total income due to non-matches is \$36.16 million, or \$35.219 if adjusted for overreporting. The non-match income effect thus amounts to 12.8% on the basis of an expected total of \$274.926 million as estimated.(14)

True Matches and Reporting Errors

A comparison of data items as they were reported for a given person to Census and to Revenue Canada reveals that inconsistencies exist at various levels of aggregation. Using RC-T data as a base for comparison, Census items are either omitted, overreported or underreported and the net result in total income may or may not be significant.

To carry out valid comparisons, income from the Census must be conceptually aligned with income reported to Revenue Canada. Thus, only those Census items were included in the comparison of total income, which are subject to provisions of the **Taxation Act**, as it applied to the 1970 taxation year. This income is referred to more precisely as "income subject to taxation" but for simplicity of exposition will be called "total income", and should not be mistaken for "total income" as it appears on the Census questionnaire.

Out of 43,886 true matches, 42,711 showed compatible total income reporting in both sources, whereas 1,158 revealed total income in their RC-T record only, 13 had zero total income in their tax return and no records, although properly matched, showed zero total income on both files. These four records, when added to 42,711 consistent records, constitute the subset of true matches with a consistent presence of total income. Consistent presence, however, does not always coincide with consistent amounts, and this aspect will have to be discussed later.

The 13 records which show zero total income on their tax return, although Census reporting shows an actual amount, will now be discussed briefly.

Isolated characteristics will be pointed out but the set is too small to permit generalizations. Five of these 13 records show wages as part of total income on their Census questionnaire, whereas the other eight records show income from self-employment or investment income.

(footnote(s) at end of text.

nécessaire de remplir une déclaration d'impôt en 1970. On a donc classé ces dossiers parmi les non-appariements justes. La valeur de ces revenus s'élève à \$11.129 millions, soit 4.0% du revenu total de l'échantillon. La différence en moins attribuable aux non-appariements est donc de \$36.16 millions, \$35.219 millions si l'on tient compte de l'exagération des revenus. L'effet des non-appariements sur le revenu s'élève donc à 12.8%, si l'on prend comme base un revenu estimatif total de \$274.926 millions(14).

Appariements justes et erreurs de déclaration

Une comparaison des données déclarées par une même personne au recensement et à Revenu Canada met en évidence certaines incompatibilités à divers niveaux de regroupement. Si l'on utilise les données de RC-T comme base de comparaison, on observe que les chiffres du recensement peuvent être omis, exagérés ou minimisés et que l'effet net d'une telle situation sur le revenu total peut être important ou non.

Pour que les comparaisons soient valables, les chiffres sur le revenu tirés du recensement doivent être conceptuellement alignés sur ceux de Revenu Canada. C'est pour cette raison que nos comparaisons ont porté uniquement sur les appariements du revenu soumis aux dispositions de la **Loi de l'impôt sur le revenu** en vigueur pendant l'année fiscale 1970. Techniquement, ce revenu correspond au "revenu soumis à l'impôt". Par souci de simplicité, nous l'appellerons "revenu total"; ce concept ne devrait toutefois pas être confondu avec le **revenu total** du questionnaire de recensement.

Des 43,886 appariements justes, 42,711 présentaient un revenu total compatible dans les deux sources de données, 1,158 avaient un revenu total dans le dossier de RC-T seulement, 13 avaient un revenu total nul dans les dossiers de l'impôt et quatre, un revenu total nul dans les deux fichiers. Ces quatre dossiers et les 42,711 autres constituent le sous-ensemble des appariements justes comportant la présence d'un revenu total compatible. Cela ne signifie pas pour autant que les sommes déclarées correspondaient toujours; cette question sera étudiée plus loin.

Nous examinerons pour le moment le cas des 13 dossiers pour lesquels les répondants ont indiqué un revenu total nul dans leur déclaration d'impôt, mais non au recensement.

Même si le nombre de ces déclarations est trop peu élevé pour qu'on en tire des généralisations, nous en isolerons néanmoins certaines caractéristiques. Dans cinq de ces 13 dossiers, les salaires font partie du revenu total; dans les huit autres, le revenu a été tiré d'un emploi autonome ou de placements.

Voir note(s) à la fin du texte.

The corresponding tax records show only "gross income from self-employment" without a corresponding "net income", or the income fields are zero. In one instance, a loss from rental income is offset by investment income thereby summing to zero total income.

The disagreement in reporting incidences can be explained in a number of ways. Income recipients may have reported Census income for 1971, since 1970 income was truly zero as revealed in their tax return. It is also possible that these are "False Matches", although initially judged true, for such a judgement is always probabilistic and never based on the absolute truth. Given that only 13 records are involved, the effect on statistical output is negligible.

The 1,158 true matches with total income exclusively on their tax return will now be examined. The magnitude of the inconsistently reported total income falls most frequently into the \$2,001 to \$5,000 income class, namely on 234 occasions. The next highest frequency occurs in the \$1,001 to \$2,000 class, where 187 records are placed. Well over one half of all records, namely 690, show total income over \$500. The average total income exclusively reported to Revenue Canada, i.e., omitted from the Census questionnaire, is \$1,695.68.

The personal characteristics of single-source respondents are of interest. Single-source respondents includes those not reporting a discernible total to Revenue Canada.

The combined number of 1,171 single-source respondents is heavily dominated by those reporting exclusively to Revenue Canada, namely 1,158. There are more females than males in this group contrary to the complete match set, which is made up of 28,344 males and 15,542 females. Marital status and age show irregular patterns.

These data are summarized in Tables 4 and 5, and population data from the 1971 Census are shown for comparative purposes in Table 2.

The 13 Census records account for a total of \$0.049 million not reflected in RC-T sources, whereas the 1,158 single-source taxation records account for \$1.963 million not reported to the Census. The average RC-T omission for this subset is \$3,769 and the average Census omission is \$1,696.

Dans les déclarations d'impôt correspondante il n'y a qu'un "revenu brut d'un travail autonome", mais aucun "revenu net", ou les champs réservés au revenu sont nuls. Dans un cas, une perte de revenu locatif a été compensée par des revenus de placements, le revenu total s'établissant ainsi à zéro.

Ces écarts peuvent s'expliquer de plusieurs façons. Les personnes qui ont eu des revenus peuvent avoir déclaré ceux de 1971, leur revenu de 1970 étant nul, comme en témoigne leur déclaration d'impôt. Il se peut également qu'il s'agisse là d'appariements erronés, le jugement qui a été porté sur la qualité de l'appariement comportant toujours une part d'aléatoire. Toutefois, comme il n'y a que 13 dossiers en jeu, l'effet sur les résultats statistiques est négligeable.

Nous étudierons maintenant les 1,158 appariements véritables où seules les déclarations d'impôt contiennent un revenu total. C'est dans la tranche de revenu de \$2,001 à \$5,000 que les erreurs de déclaration sont les plus fréquentes; nous en avons relevé 234. Vient ensuite la tranche \$1,001 à \$2,000 (187 dossiers). Dans sensiblement plus de la moitié des dossiers (690), le revenu total était supérieur à \$500. Le revenu total moyen déclaré uniquement à Revenue Canada c'est-à-dire omis dans le questionnaire du recensement - est de \$1,695.68.

Les caractéristiques personnelles des répondants présents dans une seule source sont intéressantes. Ces répondants comprennent les personnes qui n'ont pas déclaré de revenu tangible à Revenue Canada.

Le nombre total de 1,171 répondants présents dans une seule source est largement dominé par ceux qui ont fait une déclaration uniquement à Revenue Canada: 1,158. Contrairement à ce qu'on peut observer dans l'ensemble des dossiers appariés, il y a plus de personnes de sexe féminin que de personnes de sexe masculin au sein de ce groupe; au total, il y a en effet 28,344 personnes de sexe masculin et 15,542 de sexe féminin. L'état matrimonial et l'âge n'offrent pas de constantes.

Ces données sont présentées de façon sommaire dans les tableaux 4 et 5; les chiffres de population correspondants tirés du recensement de 1971 sont présentés pour fins de comparaison dans le tableau 2.

Les 13 dossiers du recensement représentent un total de \$0.049 million auxquels ne correspondent aucune somme dans les dossiers de RC-I; à l'inverse, les 1,158 dossiers présents uniquement dans les fichiers de Revenue Canada représentent une somme de \$1.963 million. L'omission moyenne dans les fichiers de RC-I s'établit à \$3,769; au recensement, la moyenne est de \$1,696.

The impact of having omitted total income is more pronounced on Census data than RC-T data, whereas partial or component omissions seem to have a greater impact upon Revenu Canada aggregates as will be seen in following sections.

Income Composition

While omission of total income is relatively infrequent in the matched set, component reporting is inconsistent to a greater degree. Table 9 provides a quick review.

Combining cells consistently reported in both sources, or consistently empty in both sources, a consistency score shows that old-age security ranks highest with 99.2%, investment income lowest with 79.4% whereas total income (subject to taxation) has been reported consistently in 97.3% of those cases where the match was judged to be true.

Inconsistent reporting does not tell the complete story. There are other problems and some of these must be viewed in the light of mis-filing of tax returns which results in mismatches. On the other hand, inconsistent reporting of components may frequently have little impact on total income. This type of effect can be associated with component substitution; i.e., a component was reported in both sources but under different classifications. "Old-age security" in the census may conceivably appear as "pension" in taxation files or vice versa. Similarly, self-employment income may have been interchanged. Other examples could be cited.

Components, as they appear in true matches, will now be examined for the purpose of quantifying possible omissions and substitutions, and their effect on "total income". The analysis will be confined to records with "income subject to taxation" present in both sources; i.e., true matches with a consistent reporting incidence of total income. If in addition to consistently reported total income all components are present in both sources, neither omissions nor substitutions exist, although magnitudes may have been reported differently.

The subset under review contains 42,711 records and represents a total income of \$7.746 million on RC-T accounts and \$6.080 million on Census accounts. Thus,

L'omission du revenu total a donc des conséquences plus marquées sur les chiffres du recensement que sur ceux de RC-I; à l'inverse, les omissions partielles semblent entraîner des conséquences plus graves pour les agrégats de Revenu Canada. C'est d'ailleurs ce que nous verrons dans les sections qui suivent.

Composition du revenu

Bien que l'omission du revenu total soit relativement peu fréquente dans les dossiers appariés, les incohérences sont plus nombreuses en ce qui concerne les éléments du revenu. Le tableau 9 en donne un aperçu.

Si l'on groupe les cases simultanément présentes ou absentes dans les deux sources, on observe que les pensions de sécurité de la vieillesse viennent au premier rang (99.2%) et les revenus de placements, au dernier (79.4%); le revenu total (soumis à l'impôt) a été déclaré dans 97.3% des appariements jugés justes.

L'étude des incohérences ne nous renseigne pas parfaitement sur la situation. Il existe d'autres problèmes, et certains d'entre eux doivent être abordés du point de vue des non-appariements attribuables à la non-production des déclarations d'impôt. En revanche, les incohérences dans la déclaration des éléments du revenu peuvent n'avoir qu'une incidence négligeable sur le revenu total. Ce genre d'effet peut être associé à la substitution des éléments du revenu: un élément a été déclaré dans les deux sources, mais sous des rubriques différentes. Ainsi, les sommes rangées sous le titre "sécurité de la vieillesse" au recensement peuvent être assimilées à des "pensions" dans les fichiers de l'impôt, et vice versa. De même, les salaires ainsi que le revenu d'un travail autonome peuvent avoir été donnés l'un pour l'autre. Les exemples ne manquent pas.

Nous étudierons maintenant les éléments du revenu tels qu'ils figurent dans les **appariements justes** afin de quantifier les omissions et les substitutions possibles et d'apprécier leur effet sur le "revenu total". Nous limiterons notre analyse aux dossiers dans lesquels le "revenu soumis à l'impôt" est présent dans les deux sources, c'est-à-dire aux appariements justes dans lesquels le revenu total a été déclaré de façon cohérente. Si le revenu total a été déclaré de façon cohérente et que ses divers éléments sont présents dans les deux sources, on peut dire qu'il n'y a ni omissions, ni substitutions; seuls les ordres de grandeur des sommes déclarées peuvent varier.

Le sous-groupe étudié contenait 42,711 dossiers qui correspondaient à un revenu total de \$237.746 millions dans les comptes de RC-I et de \$246.080 millions, dans ceux du recensement.

the matched but unweighted Census sample over-states income subject to taxation by \$8.334 million or 3.5% vis-à-vis RC-T. Normally, this net effect can be observed when comparing aggregate amounts from unmatched files. The size of the error may be judged acceptable and is usually attributed to sampling.

While sampling errors remain present, the difference described above is definitely attributable to reporting errors. It will be of more than just passing interest to analyse these reporting errors, and to reveal some of the offsetting fluctuations that result in the net effect.

With reference to Table 10, it should be noted that the reporting incidence of components is fully compatible in 27,440 cases, or 64.2% of the subset under discussion. These records account for \$141.316 million from Revenue Canada sources and \$142.939 million from the Census. The difference of \$1.623 million constitutes 1.1% of total income from Revenue Canada as shown in record pairs with consistently reported components.

Dividing the data set into three reliability categories shows that relative overreporting in Census records does not occur uniformly. The reliability categories have been defined as follows:

- A. A high-reliability grouping where the absolute deviation between total income from both sources does not exceed \$200 and where this deviation does not constitute more than 20% of Revenue Canada total income.
- B. A low-reliability grouping where the absolute deviation in total income between Census and Revenue Canada sources is more than \$200 and the corresponding percentage error is greater than 20%.
- C. An indeterminate grouping where a low absolute deviation constitutes a high percentage error, or where a high absolute deviation constitutes a low percentage error. This group exhausts the set and includes all records not classified "A" or "B".

Table 11 depicts reliability relationships and shows that category A for records with a consistent reporting incidence of components contains 17,244 records, whereas group B is the smallest with 4,580 records, and C contains 5,616 records.

Ainsi, il y a un écart de \$8.334 millions (3.5% entre l'échantillon apparié, mais non pondéré d recensement et les dossiers de RC-I. Normalement cet effet net peut s'observer si l'on compare des agrégats de fichiers non appariés. L'importance de l'erreur est acceptable; elle peut être attribuée à l'échantillonnage.

Bien que l'hypothèse de l'erreur d'échantillonnage doive être retenue, l'écart présent ci-dessus est manifestement imputable aux erreurs de déclaration. Il sera donc intéressant d'analyser ces erreurs de déclaration et de présenter certaines des variations qui en déterminent l'effet net.

Si l'on étudie le tableau 10, on observe que la fréquence de déclaration des éléments de revenu concorde pleinement dans 27,440 cas (64.2%). Ces dossiers représentent des sommes de \$141.316 millions dans les dossiers de Revenu Canada et de \$142.939 millions, dans ceux du recensement. L'écart (\$1.623 million) correspond donc à 1.1% du revenu total (RC-I).

Si l'on divise le groupe de données en trois catégories de fiabilité, on constate que la surdéclaration qui s'observe dans les dossiers du recensement ne se produit pas de façon uniforme. Les trois catégories de fiabilité ont été définies comme suit.

- A. Groupe à grande fiabilité où l'écart absolu entre le revenu total donné dans les deux sources ne dépasse pas \$200 et où cet écart ne représente pas plus de 20% du revenu total déclaré à Revenu Canada.
- B. Groupe à faible fiabilité où l'écart absolu entre les deux revenus totaux est de plus de \$200 et dans lequel le pourcentage d'erreur est supérieur à 20%.
- C. Groupe indéterminé où un faible écart absolu entraîne un fort pourcentage d'erreur et dans lequel un fort écart absolu entraîne un faible pourcentage d'erreur. Ce groupe exclut les dossiers de type A ou B.

Le tableau 11 fait ressortir les rapports entre les niveaux de fiabilité. On observe notamment qu'il y a dans la catégorie A, 17,244 dossiers cohérents (sans omissions, ni substitutions), alors que ce chiffre s'établit à 4,580 dossiers dans la catégorie B et 5,616 dossiers dans la catégorie C.

Group A departs from what appeared to be the norm of relative overreporting of Census total income. Census total income for this group amounts to \$84.722 million and Revenue Canada total income is \$84.857 million, an excess of \$0.135 million or 0.2% of Revenue Canada totals.

Group B, being a low-reliability category, shows a reporting difference of \$2.4 million with Census supplying the excess; the relevant totals are \$20.519 million and \$18.091 million for Census and Revenue Canada respectively, and the percentage error based on Revenue Canada totals is 13.4%.

The corresponding figures for group C are \$37.698 million and \$38.368 million with the excess of \$0.670 million going to the tax department, and representing 1.7% of total income from Revenue Canada files.

It can be stated in summary that consistent reporting incidence of components and loose agreement in "total income" conceals the fact that offsetting reporting errors affect subpopulations to a greater degree than any global figure could indicate.

Let us now review 15,271 records with inconsistent reporting patterns; i.e., those being encumbered with component omissions and component substitutions.

These records represent 35.8% of all true matches with consistently reported total income. They account for \$103.141 million total Census income and \$96.430 million total RC-T income, with Census being, as before, relatively high (see Table 11). The difference of \$6.711 million is 7.0% of revenue Canada derived total income. Row 10 in Table 11 shows further disaggregation of these data for the various reliability groups.

The relatively small number of 15,271 inconsistent records contributes the largest amount to the reporting error; although variation in errors between categories is large for consistent and inconsistent subsets.

The subset of records with inconsistent component reporting will now be further scrutinized. A few general statements are in order to highlight some of the underlying assumptions and basic characteristics pertaining to this subset.

Given that matched records have been judged true and that "total income" is present in both sources (Census and RC-T), components of the same type may or may not

Le groupe A s'écarter de ce qui semble être la norme du sur-déclaration relative du revenu total au recensement. En effet, le revenu total pour ce groupe s'élève à \$84.722 millions d'après les résultats du recensement et à \$84.857 millions, d'après Revenu Canada, ce qui représente un écart de \$0.135 million, soit 0.2% du total de Revenu Canada.

Le groupe B correspondant à la catégorie à faible fiabilité, l'écart est de \$2.4 millions: \$20.519 millions d'après les résultats du recensement et \$18.091 millions d'après Revenu Canada, ce qui correspond à un pourcentage d'erreur de 13.4%.

Les chiffres correspondants pour le groupe C sont de \$37.698 millions et \$38.368 millions, ce qui représente un excédent de \$0.670 million en faveur de Revenu Canada (1.7% du revenu total).

En résumé, la cohérence des éléments déclarés et la concordance du "revenu total" masquent le fait que certaines erreurs de déclaration s'annulant mutuellement touchent bien plus certaines sous-populations que les chiffres totaux ne l'indiquent.

Examinons maintenant les 15,271 dossiers où il y a eu incohérence, c'est-à-dire où certains éléments du revenu ont été omis ou substitués.

Ces dossiers représentent 35.8% de l'ensemble des appariements justes dans lesquels le revenu total a été correctement déclaré. Leur valeur s'élève à \$103.141 millions au recensement et \$96.430 millions dans les dossiers de RC-I, les chiffres du recensement étant encore une fois relativement élevés (voir tableau 11). L'écart, \$6.711 millions, équivaut à 7.0% du revenu total de Revenu Canada. Dans la ligne 10 du tableau 11, ces données sont ventilées en fonction des divers groupes de fiabilité.

Ainsi, c'est à un nombre relativement peu élevé de dossiers incohérents (15,271) qu'on doit imputer la majeure partie des erreurs de déclaration; le taux de variation des erreurs d'une catégorie à l'autre n'en demeure pas moins important.

Nous examinerons maintenant plus à fond le sous-ensemble constitué par les dossiers dans lesquels certains éléments du revenu n'ont pas été déclarés de façon uniforme. Nous donnerons auparavant un aperçu des hypothèses sous-jacentes et des caractéristiques fondamentales de ce sous-ensemble.

Supposons que des dossiers appariés l'ont été correctement et que le "revenu total" est donné dans les deux sources (recensement et RC-I), mais que les éléments du revenu peuvent avoir été

have been reported in both sources. If all equivalent components, not more, not less, have been reported in both sources, the reporting incidence is consistent and the record is not subject to the present analysis. If some or all components fail to be conceptually identical in the two sources for any given record, two possibilities arise:

- (a) the component may have been omitted in one source;
- (b) the component may have been reported under a different category heading; i.e., a substitution by way of misclassification has taken place.

In the first instance, the effect on total income would normally be larger than in the second, where total income is only affected if the substituted component also differed in magnitude.

For any given individual, if a certain component is present in the primary file (Census) without being offset by another unpaired component in the secondary file (RC-T), an omission in the secondary file must be assumed. Conversely, omissions in the primary file can be established. Similar reasoning can be applied for two, three and more omissions.

Substitution through misclassification arises when one or more components occur exclusively in the primary data string of a record and when the same number of components appears in the secondary record string, but under different classifiers. Where the number of "unpaired" components differs between primary and secondary record strings, the excess in one file origin becomes an omission in the other file origin.

The resulting classification scheme in terms of omissions and substitutions for reliability groups has been summarized in Tables 12 to 14. Records in reliability category B can be expected to contribute the largest share to the reporting discrepancies, whenever omissions are involved. The magnitude of the omissions will be reflected as a shortfall in total income and may or may not be reinforced by recall deficiencies in otherwise consistently reported income components.

Whenever substitutions dominate a set of records, it is difficult to state *a-priori* which reliability category will contribute the largest amount to total income discrepancies. However, category B remains the leading contributor to total income discrepancies under any classification scheme. A

donnés ou non dans les deux sources. Si tous les éléments déclarés dans les deux sources sont parfaitement identiques, on dit des déclarations qu'elles sont cohérentes, et le dossier n'est donc pas soumis à l'analyse qui nous intéresse. Si certains éléments ne sont pas conceptuellement identiques dans les deux sources, deux possibilités s'offrent à nous:

- a) l'élément peut avoir été omis dans une source;
- b) l'élément peut avoir été rangé dans une autre catégorie; il y a alors substitution résultant d'une erreur de classement.

L'effet de l'omission sur le revenu total est normalement plus grand que celui de la substitution, le revenu total n'étant touché que si les éléments en question diffèrent.

Pour tout dossier donné, si un certain élément est présent dans le fichier primaire (recensement), mais qu'il ne correspond à aucun autre élément non apparié du fichier secondaire (RC-T), on doit poser qu'il y a eu omission dans le fichier secondaire. Il y a omission dans le fichier primaire si le phénomène inverse se observe. Le même raisonnement peut s'appliquer à plusieurs omissions.

Il y a substitution résultant d'une erreur de classement quand un ou plusieurs éléments paraissent dans les données primaires et qu'un nombre correspondant d'éléments figure dans les données secondaires, mais sous des appellations différentes. Les éléments d'un dossier qui n'ont pas leur pendant dans l'autre correspondent à des omissions.

Le classement des omissions et des substitutions par groupes de fiabilité est présenté dans les tableaux 12 à 14. Les divergences de déclaration imputables aux omissions sont le plus souvent associées à des dossiers de la catégorie de fiabilité B. L'ordre de grandeur des omissions prend la forme d'un déficit dans le revenu total et peut ou non être aggravé par les oublis. L'égard d'autres éléments de revenu déclarés dans les deux sources.

Si les cas de substitution dominent un ensemble de dossiers, il est difficile de déterminer *a-priori* la catégorie de fiabilité qui introduit plus d'écart dans le revenu total. La catégorie demeure néanmoins à cet égard la plus importante. Un simple examen de l'écart moyen par dossier (erreur d'observation moyenne) du groupe B d

lance at the income discrepancy per record (Average NSE) for group B in Tables 12 to 14 shows that "B" retains the largest average within each subset classified by incidence of omission or substitution.

The greatest single contribution to the aggregate reporting error has been made by a group of records where one component on the census record had no counterpart on the tax return. Table 12 shows a difference in total income for this group of \$4.771 million; i.e., total income has been reported in the amount of \$4.784 million to Revenue Canada and \$9.555 million to the Census, thereby creating an average excess of \$4,287 in the census sample, this subset contains 1,113 observations.

The second greatest contribution to the reporting error of total income is also made by a group of records in category B. This group is characterized as having substituted one component and the results are shown in Table 13. There are 906 records which account for a Revenue Canada deficiency vis-à-vis Census of \$2.603 million, or \$2,873 per record.

The third-ranking group is made up of records in reliability group B and is classified as having two or three component omissions on their tax return. This group of 140 records accounts for a Revenue Canada shortfall of \$1.610 million, or \$11,500 per record. The relatively small number and the relatively large error per record suggest possible census processing errors, or the possibility of some false matches remaining in the data set.

The largest number of records where comments have been reported inconsistently appears in Table 12, reliability category A, and is made up of respondents with one component omitted in the Census. The second largest number of records originates with the same group, in reliability category C. This subset contains 2,640 records. The third ranking subset also belongs to the same group (one census component omitted), and consists of 1,139 respondents in reliability category B. In terms of membership, the fourth-ranking set consists of 1,113 respondents showing one omission in Revenue Canada records. This group was described above because it is responsible for the greatest contribution to the reporting error in aggregate dollar terms.

B dans les tableaux 12 à 14 montre que la catégorie B a la moyenne la plus élevée, quel que soit le sous-ensemble étudié.

Le facteur qui a à lui seul contribué le plus à l'erreur de déclaration globale a été l'absence dans un groupe de déclarations d'impôt d'éléments que l'on trouvait dans les dossiers du recensement. Le tableau 12 montre que la différence dans le revenu total pour ce groupe s'élève à \$4.771 millions; en d'autres termes, les revenus totaux déclarés à Revenu Canada et au recensement ont été respectivement de \$4.784 millions et \$9.555 millions, ce qui s'est traduit par un excédent de \$4,287 millions pour l'échantillon du recensement. Le sous-ensemble contient 1,113 observations.

L'erreur de déclaration du revenu total qui vient au deuxième rang est également imputable à un groupe de dossiers de la catégorie B. Dans ce cas, il y a eu substitution d'un élément du revenu (les résultats sont présentés au tableau 13). L'écart provient de 906 dossiers. Ici encore, le revenu total déclaré à Revenu Canada est inférieur aux chiffres du recensement; l'écart s'établit à \$2.603 millions, ce qui représente \$2,873 par dossier.

Au troisième rang viennent des dossiers du groupe de fiabilité B dans lesquels deux ou trois éléments ont été omis dans la déclaration d'impôt. Ce groupe de 140 dossiers représente un déficit de \$1.610 million pour Revenu Canada, soit \$11,500 par dossier. Comme il y a assez peu de dossiers et que l'erreur est relativement importante, il y a probablement erreur d'exploitation au recensement; il se peut également que le groupe contienne un certain nombre de dossiers appariés par erreur.

C'est dans le tableau 12 (catégorie de fiabilité A) que se trouve le plus grand nombre de dossiers où la déclaration de certains éléments présente des incohérences; le groupe est composé des répondants qui ont omis un élément au recensement. Le sous-ensemble qui suit appartient au même groupe, mais se présente dans la catégorie de fiabilité C; il est composé de 2,640 dossiers. Au troisième rang, viennent 2,139 dossiers du même groupe, mais de la catégorie B. Au quatrième rang, enfin, vient un groupe de 1,113 répondants qui ont commis une omission dans leur déclaration d'impôt. Ce groupe a déjà été décrit; il est en effet responsable de la plus importante erreur de déclaration en termes monétaires.

The first three largest sets described above contributed \$0.111 million, \$0.799 million and \$1.037 million respectively to the reporting error. Since these amounts are associated with Census omissions, a relative short-fall of Census income vis-à-vis Revenue Canada was observed.

The average shortfall per record is \$29, \$303, and \$485. These amounts fall within the recording capability of the Census, where amounts have been rounded to the nearest \$10.

The foregoing comments are intended to highlight the tabular material. The reader may wish to make further inferences from the data supplied in the accompanying tables.

So far, true matches have been examined for reporting consistency by attribute classes of individual records, such as reliability category, and incidence or mix of omissions and substitutions. The aggregate income effect was emphasized. This analysis will now be extended to the provincial level, where consistent and inconsistent record groups will be reviewed.

As before, the "true match" subset yields 27,440 consistent matches out of 42,711, or 64.2%. This national consistency rate ranges between provinces from a low of 62.8% in Saskatchewan to a high of 69.0% in Newfoundland. Table 10 is offered for closer study. It shows that all provinces east of Ontario have a consistency rate higher than the national average, whereas Ontario and provinces west thereof remain below the national consistency rate.

The incidence of inconsistent reporting is no indication of the effect on total income. Nationally, consistent records are associated with an error of 1.1% of total income and inconsistent records show a reporting difference of 7.0%. These errors range from 0.3% (Nova Scotia) to 5.5% (Prince Edward Island) for consistent records and from 3.3% (Ontario) to 22.7% (Saskatchewan) for inconsistent records.

The average non-sampling effect for true matches is \$59 for records with consistently reported components, and \$440 for those with components subject to omission or substitution, as observed at the national level. The average non-sampling effect ranges for consistent records between \$16 (Nova Scotia) to \$276 (Saskatchewan), and for inconsistent records between \$234 (Ontario) and \$949 (Saskatchewan).

Les trois premiers ensembles énumérés ci-dessus ont respectivement introduit des erreurs de déclaration de \$0.111 million, \$0.799 million et \$1.037 million. Ces sommes étaient associées à des omissions au recensement, les chiffres du recensement sont inférieurs à ceux du Revenu Canada.

Le déficit moyen par dossier s'établit à \$29, \$303, et \$485. Ces sommes correspondent aux possibilités de prise en compte du recensement des revenus ayant été arrondis à \$10 près.

Les commentaires présentés ci-dessus avaient pour objet de donner un aperçu des tableaux. Nous invitons néanmoins le lecteur qui aimerait étudier la question plus à fond à consulter les derniers.

Jusqu'ici, nous avons étudié la cohérence de la déclaration des appariements justes en catégorisant les dossiers (fiabilité, fréquence des omissions et des substitutions). Nous nous sommes surtout attachés aux effets de ces phénomènes sur le revenu global. Nous ferons maintenant porter notre analyse sur la répartition par province des groupes de dossiers cohérents et incohérents.

Ici encore, les sous-ensembles des appariements justes comprennent 27,440 appariements cohérents sur 42,711, soit 64.2%. Ce taux de cohérence nationale oscille entre un creux de 62.8% en Saskatchewan et un sommet de 69.0% à Terre-Neuve. Le tableau 10 mérite d'être étudié attentivement. On apprend que toutes les provinces à l'est de l'Ontario ont eu un taux de cohérence supérieur à la moyenne nationale, alors que l'Ontario et les provinces de l'ouest ont eu un taux inférieur à la moyenne nationale.

La fréquence des incohérences ne nous renseigne pas sur leur effet sur le revenu total. À l'échelle nationale, les dossiers cohérents sont en effet associés à un taux d'erreur de 1.1% du revenu total, alors que le taux d'erreur des dossiers incohérents est de 7.0%. Les taux se situent entre 0.3% (Nouvelle-Écosse) et 5.5% (Île-du-Prince-Édouard) dans le cas des dossiers cohérents et entre 3.3% (Ontario) et 22.7% (Saskatchewan), dans celui des dossiers incohérents.

En ce qui concerne les appariements justes, l'effet d'observation moyen est de \$59 dans le cas des dossiers où les éléments du revenu ont été déclarés de façon uniforme et de \$440, dans celui dont certains éléments ont été omis ou substitués. À l'échelle nationale toujours, l'effet d'observation moyen se situe de \$16 (Nouvelle-Écosse) à \$276 (Saskatchewan) dans le cas des dossiers cohérents et de \$234 (Ontario) à \$949 (Saskatchewan), dans celui des dossiers incohérents.

Ranking all 10 provinces, while including the Territories with British Columbia, the only consistent picture which emerges is that of Saskatchewan, and it is consistently inadequate. This province shows weaknesses in every respect; i.e., its consistency rate is lowest (worst) and its average reporting error for consistent and inconsistent records is highest (worst). The percentage effect on total income is also highest (worst) for records with consistent as well as with inconsistent component reporting. In other words, Saskatchewan occupies the most inferior position "10" in all classifying subsets when consistency of component reporting and associated income effects are examined.

The results for all provinces in terms of consistency rates and average reporting errors for consistent and inconsistent records are summarized in the following paragraphs.

Newfoundland ranks "one" in terms of consistency rates, "three" for average errors on consistent and inconsistent records.

Prince Edward Island ranks "three" with respect to its consistency rate, but average non-sampling effects rank "eight" and "even" for consistent and inconsistent records respectively.

Nova Scotia ranks "four" in terms of its consistency rate, occupies top spot "one" in terms of average errors on consistent records, but drops to "eight" in terms of average errors on inconsistent records.

New Brunswick ranks "two", "four" and "two" for consistency incidence, and average errors on consistent and inconsistent sets respectively.

Quebec occupies the midrange of ranks in all three categories in the order stated above, namely ranks of "five", "six", and "four" with respect to consistency incidence, and average non-sampling effects for consistent and inconsistent records.

Ontario shows a mix of ranks, namely "three", "five", and "one" for the categories mentioned above.

Manitoba fluctuates less than Ontario. It is ranked "six", "two", and "six" for the three categories under review in the order stated above.

Saskatchewan, as stated before, is consistently inadequate, it ranks "10" for all three categories.

Si l'on groupe les Territoires et la Colombie-Britannique et qu'on attribue un rang à chacune des provinces, le seul phénomène qui offre une certaine cohérence est la place occupée par la Saskatchewan. Cette province occupe en effet le dernier rang à tous égards. La Saskatchewan a le taux de cohérence le plus faible (le pire); c'est également dans cette province que l'erreur d'observation moyenne, aussi bien pour les dossiers cohérents que pour les dossiers incohérents, est la plus élevée (la pire). C'est également en Saskatchewan que l'effet sur le revenu total des dossiers cohérents et incohérents est le plus élevé (le pire). En d'autres termes, la Saskatchewan occupe le 10^e rang dans chacun des modes de classement en ce qui concerne l'uniformité de déclaration des composantes et les effets sur le revenu qui y sont associés.

Dans les paragraphes qui suivent, nous examinerons brièvement les résultats obtenus par chacune des provinces à l'égard des taux de cohérence et de l'erreur de déclaration moyenne des dossiers cohérents et incohérents.

Terre-Neuve vient au 1^{er} rang en ce qui concerne les taux de cohérence, et au 3^e pour ce qui est de l'erreur moyenne des dossiers cohérents et incohérents.

L'île-du-Prince-Édouard occupe le 3^e rang en ce qui concerne le taux de cohérence mais le 8^e et le 7^e pour ce qui est de l'effet d'observation moyen des dossiers cohérents et incohérents.

La Nouvelle-Écosse, pour sa part, vient au 4^e rang pour ce qui est du taux de cohérence, occupe le 1^{er} rang au chapitre de l'erreur d'observation moyenne des dossiers cohérents, mais passe au 8^e rang en ce qui concerne l'erreur d'observation moyenne des dossiers incohérents.

Le Nouveau-Brunswick occupe respectivement le 2^e, le 4^e et le 2^e rang.

Le Québec occupe une position moyenne dans chacune des trois catégories; il occupe en effet le 5^e, le 6^e et le 4^e rang en ce qui concerne le taux de cohérence et l'effet d'observation moyen des dossiers cohérents et des dossiers incohérents.

L'Ontario occupe diverses positions; il se classe respectivement au 9^e, au 5^e et au 1^{er} rang.

Le Manitoba occupe des positions moins extrêmes que l'Ontario. Il vient au 6^e, au 2^e et au 6^e rang.

La Saskatchewan, comme nous l'avons vu, occupe dans les trois cas la 10^e position.

Alberta falls between Saskatchewan and British Columbia, in terms of non-sampling errors. It is ranked "seven", "seven", "nine", for consistency rate, and average non-sampling income effect for consistent and inconsistent records respectively.

British Columbia, including the Territories, is second-lowest in the overall assessment. The respective rank orders are "eight", "nine", and "five".

It should be recalled that the foregoing analysis has been restricted to matched records, provided these records had been judged true. It is conceivable that a larger success rate may result in lower consistency rates. It is also possible that lower taxfiling incidences in the provinces east of Ontario may lead to more consistent data whenever a tax return has been filed and a match was brought about. In other words, a tradeoff in quantity versus quality may exist at the collection stage, where data collection relates to the filing of tax returns.

To follow this line of reasoning, taxfiler rates have been calculated and are shown in Table 15. They have been expressed as a percentage of the 1971 adult Census population (15 years and over), and the number of tax returns filed by early spring of 1971. It can be readily observed that all provinces east of Ontario remain below the national rate of 60.2%, whereas all provinces west of Quebec, except Saskatchewan, are above the national rate.

Since matching a record successfully presupposes the filing of a tax return, one may hypothesize that the match rate is correlated with the taxfiler rate. The match rate is defined as the ratio of true matches out of the estimated tax universe expressed in per cent. The propensity to file a tax return is based exclusively on Census income information, and dependency relationships within families.

Match rates are also shown in Table 15. The national rate of 91.5% is exceeded by all provinces west of Quebec, and by Nova Scotia in Eastern Canada. Ranking all provinces by their taxfiler rate and by their match rate reveals that these two rates are correlated. The statement is based on a Spearman rank correlation coefficient of 0.70 which is significant at the 5% level.

Instead of using the match rate, as previously defined, a success rate was calculated. This rate is made up of the sum of

En ce qui concerne l'erreur d'observation l'Alberta se place entre la Saskatchewan et la Colombie-Britannique. Elle vient au 7e rang pour ce qui est du taux de cohérence et respectivement au 7e et au 9e rang en ce qui concerne l'effet d'observation moyen des dossiers cohérents et incohérents.

La Colombie-Britannique (Territoires compris) occupe globalement l'avant-dernier rang. Elle est classée respectivement 8e, 9e et 5e.

Il convient de rappeler que cette analyse a été limitée aux dossiers qui ont fait l'objet d'un appariement juste. Il est raisonnable de croire que plus le taux de réussite augmente, plus le taux de cohérence diminue. Il se peut également que la faiblesse relative du taux de déclaration à l'impôt dans les provinces à l'est de l'Ontario fasse que les données soient plus cohérentes quand une déclaration d'impôt a été produite et qu'il y a eu appariement. En d'autres termes, il se peut qu'on doive attacher plus d'importance à la qualité qu'à la quantité de données au moment de la collecte, particulièrement si les chiffres recueillis ont rapport à la production de déclarations d'impôt.

C'est pour cette raison que nous avons calculé les taux de déclaration à l'impôt; ces chiffres sont présentés au tableau 15. Ils sont exprimés en pourcentage de la population adulte au recensement de 1971 (15 ans et plus) et du nombre de déclarations d'impôt produites au début du printemps 1971. Il est facile de voir que toutes les provinces à l'est de l'Ontario demeurent sous le taux national de 60.2%, alors que les provinces à l'ouest du Québec - exception faite de la Saskatchewan - se placent au-dessus du taux national.

Comme l'appariement ne peut se faire que sur une déclaration d'impôt a été produite, on peut supposer que le taux d'appariement est lié au taux de déclaration à l'impôt. Par définition, le taux d'appariement équivaut au rapport entre le nombre d'appariements justes et l'univers fiscal estimatif; il est exprimé en pourcentage. La tendance à produire une déclaration d'impôt est mesurée uniquement à partir des chiffres du recensement sur le revenu et des liens de parenté des personnes au sein des familles.

Les taux d'appariement sont également présentés au tableau 15. Le taux national de 91.5% est dépassé par toutes les provinces à l'ouest du Québec et par la Nouvelle-Écosse. Si l'on classe les provinces en fonction du taux de déclaration à l'impôt et du taux d'appariement, on constate qu'il y a un rapport entre les deux. Cette observation s'appuie sur le coefficient de corrélation par rangs Spearman (0.70, significatif à 5%).

Plutôt que d'utiliser le taux d'appariement nous avons calculé un taux de réussite. Pour établir le taux, nous avons fait la somme de

all true matches and all true non-matches as a percentage of all adults in the sample. It can also be viewed as the complement to the failure rate which is expressed in terms of all "false non-matches" out of all adults in the sample. It should be recalled that false non-matches are the only unresolved failures after "false matches" have been converted into non-matches and then judged "true or false" with respect to the non-match decision.

The provincial success rate is highly correlated with the provincial match rate. Rank correlation of these two rates is 0.86, and is statistically significant at the 5% level. The rank correlation coefficient between success rate and taxfiler rate, however, is no longer significant. Lack of significance compared to significance of the first two rates may indicate the possibility of improving match rates by upgrading taxfiler coverage, whereas the validity of non-match decisions remains independent thereof.

Consistent reporting of components on true matches is inversely correlated with the provincial match rate, also judged by the Spearman rank correlation coefficient. The coefficient is - 0.78, which is significant at the 5% level. This negative correlation could imply that higher taxfiling activity is accompanied by a greater incidence of omissions or substitutions. It could also mean that in high-taxfiler areas, a higher percentage of taxfilers is completing Census questionnaires without resorting to tax return comparisons, thereby increasing the incidence of inconsistently reported items.

The discussion of non-sampling effects in terms of consistent component reporting, impact on total income, and provincial variation will now shift to specific components. Selected components will be discussed in terms of likely substitutability as a result of misclassification by the respondent.

In interview surveys, and probably even more frequently in self-enumeration surveys, such as the Census, misclassification of income components arises from lack of perception by the respondent. Either instructions are not read, are misinterpreted, or items are entered into questionnaires with preconceived ideas. Since the reason for component collection must often be sought in the desire to get more reliable totals through avoidance of omissions, the level of individual components may be of limited interest. Nevertheless, reconciliations by components are attempted later on. Moreover,

appariements et des non-appariements justes et nous l'avons exprimée en pourcentage du nombre d'adultes dans l'échantillon. Il correspond d'une certaine façon au complément du taux d'échec qui équivaut à la proportion "non-appariements erronés/adultes dans l'échantillon". Il convient de se rappeler que les non-appariements erronés ne représentent que les échecs non résolus après que les appariements erronés ont été convertis en non-appariements, puis jugés justes ou erronés.

Le taux de réussite provinciale est intimement lié au taux d'appariement provincial. Le facteur de corrélation par la méthode des rangs de ces deux taux est de 0.86; son niveau de signification statistique est de 5%. Le coefficient de corrélation entre le taux de réussite et le taux de déclaration à l'impôt, en revanche, perd toute signification; cela signifie peut-être qu'on pourrait accroître les taux d'appariement en améliorant l'observation des contribuables, la validité des décisions relatives aux non-appariements demeurant insensible à la qualité de l'observation.

Comme le montre le coefficient de corrélation Spearman, il y a un rapport inverse entre la cohérence de la déclaration des éléments du revenu dans les appariements justes et le taux d'appariement provincial. Le coefficient, dans ce cas, est de - 0.78 (il est significatif à 5%). Ce coefficient de corrélation négatif signifie peut-être que les forts taux de déclaration à l'impôt sont liés à une plus grande fréquence d'omissions ou de substitutions. Il se peut également que, dans les régions où le taux de déclaration à l'impôt est élevé, une plus forte proportion des contribuables remplissent leur questionnaire de recensement sans consulter leur brouillon d'impôt, ce qui accroît d'autant les risques d'incohérence.

Nous analyserons maintenant les effets d'observation en fonction de certains éléments du revenu. Nous aborderons notamment la question sous l'angle des possibilités de substitution attribuables à des erreurs de classement des répondants.

Dans les enquêtes par interview - et plus souvent encore dans les enquêtes par autodénombrement telles que le recensement - les erreurs de classement des éléments du revenu sont le plus souvent imputables aux fautes d'interprétation des répondants. Il se peut par exemple que les répondants ne lisent pas les instructions, qu'ils les comprennent mal ou qu'ils inscrivent leurs réponses en ayant des idées préconçues. Comme on recueille le plus souvent les données sur le revenu par éléments afin d'obtenir des chiffres plus fiables en évitant des omissions, les éléments n'offrent parfois qu'un intérêt limité. Il arrive néanmoins qu'on procède à des rapproche-

major income sources are determined on the basis of component reporting, and various statistics are produced on the strength of these components.

As was mentioned earlier, the reliability of components is often determined on the basis of observable differences between data sources. Such differences, however, are net of offsetting errors; e.g., one may observe a net difference in farm income which may indicate that farm income is too small relative to some other source. However, some **farm income** may have been reported under other categories, thereby creating a deficiency, whereas other **non-farm components** may have slipped into the farm slot, thereby reducing this deficiency.

Employment income components have been misclassified in the past, although a precise quantification was always difficult to ascertain. It will now be attempted to quantify the misclassification effect. Substitutions because of misclassification occur between wages and salaries, non-farm income from self-employment, and farm income from self-employment. Substitutions may take place in any combination, but for analytical purposes it is assumed that any substitution within the employment-income subset involves only two components for any record.

A record may reveal wages and salaries exclusively on the Census questionnaire and farm income exclusively on the RC-T file. It is assumed under these circumstances that wages on the Census questionnaire have been substituted for farm income. Similarly, another record may show **farm income** (from self-employment) exclusively on the tax return and **non-farm income** from self-employment exclusively on the Census questionnaire. Substitution of non-farm for farm income on the Census questionnaire is assumed under these circumstances.

If a given record were to have wages on the tax return but no other employment income, and farm income and non-farm income from self-employment on the Census questionnaire, a substitution of farm income for wages as well as a substitution of non-farm income from self-employment for wages would be counted, where in fact one substitution and one omission could exist. However, the likelihood of this occurrence is remote. Only 2.5% (22/870) of all records with farm net income as a major source show non-farm self-employment income as a secondary source. For non-farm self-employment income as a major source only 4.2% (69/1,657) of all records have farm income as a minor source. While these figures do not supply

ments par élément. De plus, les chiffres sur les principales sources de revenu et diverses autres statistiques sont établis en fonction de ces éléments.

Comme nous l'avons déjà vu, la fiabilité des éléments du revenu est souvent déterminée en fonction des différences qui peuvent s'observer entre les sources de données. Ces écarts, toutefois, ne sont pas exempts d'erreurs. Ainsi, on pourra observer un écart net au titre du revenu agricole qui signifiera que le revenu agricole est trop peu élevé par rapport aux autres sources. Toutefois, certains **revenus agricoles** peuvent avoir été rangés dans d'autres catégories et certains **revenus non agricoles** peuvent avoir été assimilés par erreur à des revenus agricoles. L'écart créé par le premier groupe d'erreurs étant réduit d'autant.

Les éléments du revenu de l'emploi ont souvent été mal classés dans le passé, bien qu'il soit difficile d'établir dans quelle proportion. Nous allons ici tenter de quantifier l'effet de ces erreurs de classement. Les substitutions attribuables aux erreurs de classement touchent la rémunération, le revenu non agricole tiré d'un emploi autonome et le revenu agricole tiré d'un emploi autonome. Les substitutions peuvent prendre plusieurs formes; pour les fins de l'analyse, toutefois, nous supposons qu'elles ne touchent que deux éléments du revenu dans un même dossier.

Un dossier peut faire état de rémunération dans le questionnaire du recensement et d'un revenu agricole dans le fichier de RC-I. Dans ces circonstances, on suppose que les rémunérations déclarées au recensement ont en fait été substituées à un revenu agricole. De même, il peut arriver qu'un contribuable ait indiqué un **revenu agricole** (tiré d'un emploi autonome) dans sa déclaration d'impôt, mais qu'il ait indiqué un **revenu non agricole** tiré d'un emploi autonome dans le questionnaire du recensement. Dans ces circonstances, on suppose que le revenu non agricole a été substitué au revenu agricole dans le questionnaire du recensement.

Si un dossier contient des rémunérations dans la déclaration d'impôt, mais aucun autre revenu de l'emploi, et un revenu agricole ainsi qu'un revenu non agricole tiré d'un emploi autonome dans le questionnaire du recensement, on établit que les rémunérations ont été remplacées par un revenu agricole et un revenu non agricole tiré d'un travail autonome, ce qui équivaut à deux substitutions, là où il n'y avait peut-être qu'une substitution et une omission. Toutefois, il est peu probable qu'une telle situation se produise. On observe en effet que le revenu non agricole tiré d'un travail autonome ne constitue la deuxième source de revenu des personnes dont la principale source de revenu est le revenu agricole net que dans 2.5% des cas (22/870). Par ailleurs, seulement 4.2% des répondants dont l

conclusive evidence, they are sufficient to alleviate any fears that the data in support of the substitution hypothesis are heavily afflicted with double counting.

Table 16 summarizes the incidence of net income from farming, non-farm income from self-employment, and wages reported in one source without its equivalent counterpart in the other source, but with another employment-income component in the secondary source reported exclusively therein. The upper left three-by-three sections of Table 16 serves to illustrate this situation. The upper right and lower left three-by-three sections contain supplementary information and help to convey a sense of proportion with respect to the incidence of substitution.

To assist in reading the table, a few of the depicted relationships will be spelled out. There are 54 records with non-farm self-employment reported exclusively to RC-T but showing farm net income on their census return. There are 56 records with wages and salaries on their tax return, but their census return shows farm net income as a single-source item. It is assumed in each instance that RC-T information is correct and Census information is incorrect. This assumption is based on recall phenomena which are sound for RC-T reporting because documentary evidence is required. There is no such provision of sanctions which is apt to reduce reporting errors on tax returns. Conversely, recall on Census questionnaires is relatively poor because documentary evidence is not required, sanctions are non-existent, and the elapsed time since the reporting took place is greater than for RC-T reporting.

The supplementary information shows that on Census questionnaires with farm income and non-farm self-employment income on the census questionnaire and on the RC-T file, thereby prohibiting a substitution of farm income for non-farm self-employment income. Alternatively, they could also have had zero returns for non-farm self-employment income in both sources. This type of occurrence permits the same inference of non-substitution.

A similar situation is depicted in the lower-left quadrant where 294 RC-T farm income records show non-farm self-employment occur either in both sources or not at

principale source de revenu est un revenu non agricole tiré d'un travail autonome (69/1,657) tirent également un revenu de l'activité agricole. Bien que ces chiffres ne nous permettent pas de tirer des conclusions définitives, ils lèvent néanmoins les doutes selon lesquels les données qui appuient l'hypothèse de substitution risquent de faire l'objet de doubles comptes.

Le tableau 16 présente de façon sommaire les cas où un revenu agricole net, un revenu non agricole tiré d'un travail autonome et des rémunérations ont été déclarés dans une source, sans que leur contrepartie figure dans l'autre source, mais où un autre revenu d'un emploi est donné uniquement dans la deuxième source. Le cadre supérieur gauche du tableau 16 illustre bien la situation. Le cadre supérieur droit et le cadre inférieur gauche du tableau contiennent des renseignements supplémentaires qui mettent dans une meilleure perspective la question de la fréquence des substitutions.

Pour faciliter la lecture du tableau, nous décrivons certains des rapports qui y sont mis en évidence. Il y a 54 dossiers dans lesquels le répondant a déclaré un revenu non agricole tiré d'un travail autonome à RC-T, mais un revenu agricole net au recensement. Par ailleurs, on observe que 56 répondants ont déclaré à l'impôt qu'ils touchaient une rémunération, alors que dans leur questionnaire de recensement, ils ont indiqué que le revenu agricole constituait leur seule source de revenu. Dans chaque cas, nous avons posé que les renseignements de RC-T étaient exacts et que ceux du recensement étaient erronés. Cette décision s'appuie sur le fait que les contribuables doivent ajouter à leur déclaration un certain nombre de pièces justificatives. De plus, le fisc a prévu des peines qui sont susceptibles de réduire le taux d'erreur. En revanche, dans une enquête mémoire comme le recensement, on n'exige pas de pièces justificatives, il n'y a pas de peine prévue et il s'est écoulé plus de temps entre le moment où le revenu a été gagné et le jour du recensement qu'entre la même époque et la période où le contribuable a produit sa déclaration d'impôt.

Les renseignements supplémentaires montrent que les 235 répondants au recensement qui ont déclaré un revenu agricole avaient également déclaré au recensement et à l'impôt un revenu non agricole tiré d'un travail autonome; il ne pouvait donc pas y avoir substitution dans ces cas. Par ailleurs, si les répondants avaient indiqué dans les deux sources qu'ils n'avaient pas tiré d'un travail autonome un revenu non agricole, on aurait pu en tirer les mêmes conclusions.

La situation est la même dans le cadre inférieur gauche; si l'on prend par exemple les 294 dossiers dans lesquels un revenu agricole a été déclaré à l'impôt et qu'on constate qu'un revenu

all, thereby precluding an assumption of component substitution. The remaining data in Table 16 should be interpreted in a similar fashion.

Returning to the "presumption of innocence" for RC-T records; i.e., they are assumed to be free from fault until proven otherwise, one can translate the entries in the upper left quadrant of Table 16 into corrections which could be applied to the Census. The 54 and 56 records in the first row of Table 16 would be applied as a negative correction to Census farm income. Although shown as farm income on the Census, their true component membership is determined by RC-T information, and it places them outside the farm universe. The same records would also constitute a positive correction to Census non-farm self-employment income (54), and to Census wages and salaries (56). Census failed to show entries for these cells, but RC-T data indicate that these income components should have been assigned accordingly.

Negative and positive corrections must balance for the complete data set. Thus, Census farm income should have been added to 83 records as the net result of this process. Census non-farm self-employment income should also have been assigned to an additional 89 records, whereas Census wages should have been removed in 172 instances.

While the net effect is largest for wage earners, it is based on offsetting negative and positive corrections of 661 and 489 records. Non-farm self-employment income, on the other hand, would face corrections of a similar magnitude, namely 538 negative and 627 positive ones, but the net effect is reduced to 89 cases. For farm income, the net effect is similar to that of non-farm self-employment income, but it is based on the smallest set of corrections juxtaposing 110 and 193 records for a net effect of 83.

The foregoing examples were presented in order to illustrate that net effects are not always indicative of reporting qualities. Normally, only these net effects can be observed. It is the increased power of observation attributed to a linked data set which permits a closer assesment of these reporting phenomena.

For investment income, a different form of substitution has been hypothesized, namely reporting investment income under the husband's name in one source, and under the

non agricole tiré d'un travail autonome est simultanément présent ou absent dans les deux sources, on peut en déduire qu'il n'y a pas eu substitution d'éléments. Les autres données du tableau 16 peuvent être interprétées d'une manière analogue.

Revenons à la présomption de qualité des dossiers de RC-I. Si l'on suppose qu'ils sont exacts tant que le contraire n'a pas été prouvé, on peut en déduire que les chiffres du cadre supérieur gauche du tableau 16 correspondent à des corrections qui pourraient être appliquées aux résultats du recensement. Ainsi, les 54 et 56 dossiers de la première ligne du tableau pourraient prendre la forme de corrections négatives apportées aux chiffres du recensement sur le revenu agricole. Bien qu'on les assimile au recensement à un revenu agricole, leur appartenance réelle est fixée en fonction des chiffres de RC-I, ce qui les sort de l'univers agricole. Ces dossiers pourraient également prendre la forme d'une correction positive apportée au revenu non agricole tiré d'un travail autonome (54) et aux rémunérations (56). Au recensement, il n'y a pas de chiffre dans ces cases; les données de RC-I nous apprennent néanmoins le contraire.

Au total, les corrections négatives et positives doivent s'équilibrer. Ainsi, 83 dossiers viendraient s'ajouter à ceux du revenu agricole au recensement et 89, au revenu non agricole tiré d'un travail autonome du recensement; en revanche, 172 dossiers devraient être enlevés des rémunérations.

Bien que l'effet net de cette opération touche davantage les personnes qui gagnent un revenu, les calculs reposent sur des corrections négatives et positives apportées à 661 et 489 dossiers. Le revenu non agricole tiré d'un travail autonome, pour sa part, ferait l'objet d'un nombre sensiblement égal de corrections (538 négatives et 627 positives), le nombre net de dossiers diminuant de 89. Pour ce qui est du revenu agricole, enfin, l'effet net de l'opération serait voisin, car il toucherait 83 dossiers; toutefois, il reposerait uniquement sur 110 et 193 corrections.

Nous avons donné ces exemples pour illustrer le fait que les effets nets de telles opérations ne nous renseignent pas toujours bien sur la qualité des déclarations. Habituellement, seuls ces effets nets peuvent s'observer. Seul le couplage des données nous permet d'étudier de plus près ces phénomènes.

En ce qui concerne les revenus de placements, on a supposé qu'il pouvait y avoir deux types de substitution: la déclaration des revenus de placements au nom d'un conjoint dans une source et

ife's name in the other source. This sort of shifting may apply particularly to interest from joint savings accounts and bond interest. The shifting of bond interest could occur if the security was bought by one spouse in the other spouse's name. Although the RC-T treatment is clear, namely the interest should be reported by the purchaser (donor) of the bond, and not the registered owner, Census reporting may not follow these lines.

There is insufficient evidence to support this hypothesis. Only 89 records show investment income reported to RC-T by husbands but not by their respective spouses, and reported under the wife's name to the census without any of it reported under the husband's name. There are also 174 records with the reverse reporting relationship. The net gain for Census wives thus is 85 with a corresponding net loss to the husband's column. However, these numbers are extracted from about 4,000 couples with at least one of the marriage partners reporting investment income. The net effect therefore constitutes about 2% of the total number of records under consideration. While investment income is afflicted with omissions in both files, as can be seen from Table 9, it seems to be relatively free from substitution among marriage partners. Partial substitution of investment components, however, may occur, but this phenomenon cannot be measured with the data at hand.

The substitution of income components and their reporting errors may affect aggregate income and income distributions disproportionately. To gain some insight into the income effect, matched records will now be discussed with reference to "squareables". In these tables, class membership from one source is cross-classified with class membership from the other source. If all income recipients had reported their income identically to Census and to Revenue Canada, all entries would be located along the main diagonal of such a table.

The choice of class limits could shift marginal records by one income class. Thus, agreement between sources is usually judged in terms of records on or immediately adjacent to the main diagonal.

Class size, of course, will influence the fit. If, for example, a distribution were to be restricted to three income classes, all individuals would fall on or immediately adjacent to the main diagonal. Conversely, an "infinitely" large number of income classes would leave very few records within the designated limits of "good fit".

au nom de l'autre conjoint dans la deuxième source. Cette situation se produit probablement le plus souvent dans le cas des intérêts des comptes d'épargne conjoints et des intérêts d'obligations, particulièrement si les obligations ont été achetées par un conjoint au nom de l'autre. Bien que le traitement retenu par RC-I soit clair - les intérêts doivent être déclarés par l'acheteur (le donateur) et non par le propriétaire enregistré - les répondants au recensement ne procèdent peut-être pas de cette façon.

Les preuves pour appuyer cette hypothèse manquent. En effet, il n'y a que 89 dossiers où des revenus de placements ont été déclarés à RC-I par des époux, mais non par leur épouse, alors qu'ils ont été uniquement déclarés au nom de l'épouse au recensement. Par ailleurs, la situation inverse s'observe dans 174 dossiers. Les épouses réalisent donc un gain net au recensement de 85; il y a évidemment une perte correspondante pour les époux. Il convient néanmoins de souligner que ces chiffres sont tirés de données sur près de 4,000 couples dont au moins un des conjoints a déclaré des revenus de placements. L'effet net s'établit donc à environ 2% des dossiers étudiés. Bien que les revenus de placements fassent l'objet d'omissions dans les deux fichiers (cf., tableau 9), les cas de substitution entre conjoints semblent relativement peu fréquents. Les cas de substitution partielle qui auraient pu se produire n'ont pas pu être mesurés faute de données.

La substitution des éléments du revenu et les autres erreurs de déclaration ne touchent pas nécessairement de la même façon le revenu global et les répartitions du revenu. Pour avoir une certaine ouverture sur l'effet du revenu, nous étudierons les dossiers appariés à l'aide de tableaux carrés. Dans ces tableaux, l'appartenance d'une source à une classe donnée est étudiée en fonction de l'appartenance d'une autre source à la même classe. Ainsi, si toutes les personnes qui avaient touché un revenu l'avaient déclaré de la même façon au recensement et à Revenue Canada, les données se situeraient toutes dans la diagonale principale de ce tableau.

Comme le choix des limites des classes pourrait faire passer certains dossiers marginaux d'une classe à une autre, on a posé qu'il y avait concordance quand un dossier se situait sur la diagonale principale ou immédiatement à côté.

La qualité de l'ajustement dépend évidemment de la taille des classes. Si, par exemple, on limitait la répartition à trois classes de revenu, toutes les personnes observées tomberaient sur la diagonale principale ou dans un secteur adjacent. À l'inverse, l'utilisation d'un très grand nombre de classes de revenu ne laisserait que quelques dossiers dans les limites des "bons" ajustements.

Initially, a square table with 38 income classes was produced, and "total income" as well as "wages and salaries" were cross-tabulated therein. The universe was restricted to "true matches". Total income, made conceptually compatible for both sources, showed 75.6% of all respondents within one class interval of the main diagonal and for wages the result was 77.9%.

There is no precise measure which states that a given percentage is "good" or "bad". However, some empirical evidence showed what might be "attainable". The United States Department of Health, Education and Welfare (HEW) had published square tables on wages, where matched records within one class interval of the main diagonal showed 85.4% agreement.⁽¹⁵⁾ However, the United States study was based on 18 classes rather than 38, as was the case for our data.

When our results for wages were retabulated using the same 18 class intervals that had been used by HEW, our fit improved from 77.9% to 85.0%. The percentage on the main diagonal proper was 67.1%, which was a slight improvement over the United States results, where the main diagonal proper contained 65.2% of all records.

The foregoing discussion of square tables is summarized in Table 18. Additional income components are also shown therein and the mediocre reporting quality discussed for those components earlier is supported by "square table" presentation.

The discussion of matching results will now return to the non-matched records.

Non-matched Records from the 1971 Census

Non-matches were briefly juxtaposed with matches in the introductory section to match results. It was stated that the non-match set consisted of 33,516 original non-matches and 1,779 "converted" non-matches for a total of 35,295.

It was also stated that a non-match decision could be a correct decision (true non-match), namely when the Census record to be matched belongs to a person who is not an income recipient, or whose income is relatively small, and based on sources which reduce the likelihood of filing a tax return. It was stated without further support that 4,084 non-matches were **false non-matches** (see also Table 8) whereas 31,211 records must be considered true non-matches (see also Table 7).

See footnote(s) at end of text.

À l'origine, nous avons produit un tableau carré qui comportait 38 classes de revenu et nous y avons porté le revenu total ainsi que les rémunérations. L'univers était limité aux appariements justes. En ce qui concerne le revenu total - qui avait au préalable été rendu conceptuellement compatible pour les deux sources - 75.6% des répondants se trouvaient à une classe d'intervalle de la diagonale principale; dans le cas des rémunérations, la proportion correspondante était de 77.9%.

Il n'existe pas de mesure précise qui nous permette de dire si un pourcentage donné est "bon" ou "mauvais". Toutefois, certains résultats empiriques nous renseignent sur les résultats qui pourraient être atteints. Ainsi, le ministère américain de la Santé, de l'Éducation et du Bien-être a publié des tableaux carrés sur les rémunérations dans lesquels 85.4% des dossiers appariés tombaient à une classe d'intervalle près de la diagonale principale (15). L'étude américaine reposait néanmoins sur 18 classes, et non sur 38.

Après que nous ayons eu présenté nos résultats sur les rémunérations en fonction des 18 classes utilisées dans l'étude américaine, l'ajustement est passé de 77.9% à 85.0%. 67.1% des dossiers se trouvaient sur la diagonale principale, ce qui est légèrement mieux que dans l'étude américaine (65.2%).

Ces résultats sont présentés de façon sommaire au tableau 18. Le tableau contient également des chiffres sur d'autres éléments du revenu; la piètre qualité de déclaration relative à ces éléments est d'ailleurs bien mise en évidence par la présentation du tableau.

Nous reviendrons maintenant aux résultats de l'appariement des dossiers non appariés.

Dossiers non appariés du recensement de 1971

Nous avons brièvement comparé les non-appariements aux appariements dans l'introduction de la section des résultats de l'appariement. Nous avons notamment vu que les 35,295 non-appariements comprenaient 33,516 non-appariements d'origine et 1,779 appariements rejetés.

Nous avons également vu que les décisions de non-appariement pouvaient être justes quand le dossier du recensement appartenait à une personne qui n'avait pas eu de revenu ou dont le revenu était relativement peu élevé et provenait probablement de sources moins susceptibles de fausser l'objet d'une déclaration d'impôt. Nous avons également affirmé que 4,084 non-appariements étaient des **non-appariements erronés** (voir également le tableau 8), alors que 31,211 dossiers devaient être assimilés à des non-appariements justes (tableau 7).

Voir note(s) à la fin du texte.

The majority of **true non-matches**, namely 19,939 have no **income subject to taxation** reported on their Census questionnaire; i.e., income which could be taxed if sufficient amounts had been received. A person with such a record could have received family allowances, veterans' pensions, or workmen's compensation, for example, which are **not subject to taxation** in 1970.

The "converted" non-matches were originally classified as "false matches"; i.e., a given Census record had been matched erroneously with a tax record of similar characteristics. It was then decided during post-match edit that these records constituted an invalid combination, and that the match should be disbanded by having the Census record revert to its original form and become part of the non-match set.

Given that it was retained as a non-match, the truthfulness of this decision can be questioned. Out of 1,779 "converted" non-matches (formerly false matches), 817 were judged false again, whereas 962 were classified as true non-matches. In other words, the decision to convert "false matches" into "non-matches" resulted in 45.9% of these records to have their match status classified correctly, whereas 54.1% remained problem cases.

It should be recalled that the classification of non-matches as "true" or "false" is based on the degree of likelihood with which an individual represented by such a record can be expected to file a tax return. It must be assumed that the Census information is correct, for this information forms the basis for judging the propensity to file a tax return.

One should also remember that Census information is subject to omissions and substitutions. Consequently, non-matched Census records can also be expected to contain substitutions and omissions. Such omissions, if rectified, would increase the number of true non-matches so classified, whereas the general tendency to overreport income on the Census would overstate the number of false matches initially. Consequently, judging the truthfulness of non-match decisions on the basis of income reported to the Census will result in minor distortions with little net effect.

The propensity to file a tax return is judged against general taxfiling criteria, such as the size of income, the dependency status, the income of dependents, and age, which may entitle the recipient to an age exemption. Allowance cannot be made for special deductions due to pension plan contri-

La majorité des **non-appariements justes** (19,939) n'ont aucun **revenu soumis à l'impôt** dans le questionnaire du recensement (il s'agit ici de revenus qui auraient pu être imposés s'ils avaient été suffisamment élevés). Ces personnes peuvent voir reçu des allocations familiales, une pension d'ancien combattant ou des indemnités pour accident du travail qui n'étaient pas **soumises à l'impôt** en 1970.

Les non-appariements "modifiés" étaient à l'origine rangés dans les "appariements erronés"; on supposait qu'un dossier du recensement avait été apparié par erreur à une déclaration d'impôt qui avait des caractéristiques voisines. On a ensuite décidé pendant le contrôle qui a suivi l'appariement que ces dossiers formaient une combinaison invalide, que l'appariement devait être annulé et que le dossier du recensement devait revenir à son point d'origine et être réintégré à l'ensemble des dossiers non appariés.

Compte tenu du fait qu'on posait alors qu'il y avait non-appariement, la justesse de cette décision peut être mise en doute. Des 1,779 non-appariements modifiés (appariements qualifiés auparavant d'erronés), 817 ont été jugés inexacts et 962 ont été rangés dans la catégorie des non-appariements justes. En d'autres termes, suite à la décision de modifier les appariements erronés en non-appariements, 54.1% des dossiers visés ont reçu un bon statut d'appariement, alors que 45.9% faisaient toujours problème.

Il convient de rappeler que le classement des non-appariements en non-appariements justes ou erronés est fondé sur la probabilité selon laquelle la personne représentée par un dossier a des chances d'avoir produit une déclaration d'impôt. Il faut donc supposer que les données du recensement sont exactes, car c'est sur elles qu'on s'appuie pour déterminer la tendance à produire une déclaration d'impôt.

Il faut également se rappeler que les données du recensement font l'objet d'omissions et de substitutions. Les dossiers non appariés ne font pas exception. Si ces omissions étaient corrigées, on accroîtrait le nombre des non-appariements erronés, et la tendance générale à déclarer au recensement des revenus trop élevés provoquerait une surévaluation du nombre des appariements erronés. Il en résulte donc que la détermination de la qualité des décisions relatives aux non-appariements en fonction du revenu déclaré au recensement ne donne lieu qu'à de légères distortions sans effet net marqué.

La tendance à produire une déclaration d'impôt est jugée en fonction des critères généraux de production d'une déclaration, de la taille du revenu, du statut de personne à charge, du revenu des personnes à charge et de l'âge (qui peut donner droit à une exemption en raison d'âge). Il n'est pas possible de tenir compte des déductions

butions, medical deductions, or alimony paid. The likelihood of filing a tax return is further modified by institutional constraints due to provisions for withholding taxes on wages. Thus, *ceteris paribus* wages, or an appreciable wage component, may increase the likelihood of a non-match being false.

The impact of components on the classification of non-matches was derived from major-source-of-income determination. Differentiating features were wages versus self-employment income, and non-employment income.

True and false non-matches will now be discussed in terms of their characteristics. The true non-matches are of relatively little interest, except that they constitute a subset which will always leave a data gap. By definition, a true non-match does not lend itself to any remedial action which would result in the missing information being **added**. A false non-match, on the other hand, may be subject to remedial action. It could consist of improved data collection, revised matching methods, or amelioration with the help of synthetic linkage.

Whenever the decision to declare a non-match happens to be a false one, the following questions should be asked: are these non-matches false because their tax records cannot be found although they exist?, or are they false because the tax records do not exist, although tax returns should have been filed?

The income data associated with non-matches can be summarized as follows: Out of 31,211 **true non-matches**, 19,939 (63.9%) have no income subject to taxation, whereas 11,272 (36.1%) have some income, but the corresponding recipient must be considered non-taxable. Considering the subset of 11,272 non-matches with some income subject to taxation, 96.5% (10,878 records) have reported income under \$2,500, and 5,929 of these have reported income under \$1,000.

False non-matches have no members at the low end of the income distribution. The distributional impact in terms of membership and income can be gleaned from Table 19, where the potential universe consists of all matches and non-matches. The non-match effect is broken down for true and false categories.

While a 19.0% shortfall in membership results in a 4.0% shortage of income for the universe due to **true non-matches**, the effect

de tenir compte des déductions spéciales telles que les contributions à un régime de retraite les déductions pour frais médicaux ou les pensions alimentaires versées. La tendance à produire une déclaration d'impôt est également liée aux contraintes administratives suscitées par les retenues d'impôt à la source. Ainsi, toutes choses étant égales par ailleurs, les rémunérations - ou un élément "rémunérations" important - peuvent accroître le risque qu'un non-appariement soit erroné.

L'incidence des éléments du revenu sur le classement des non-appariements a été évaluée en fonction de la principale source de revenu. Les principaux facteurs de différenciation étaient les rémunérations versus le revenu d'un travail autonome et le revenu hors-travail.

Nous étudierons maintenant les caractéristiques des non-appariements justes et erronés. Les non-appariements justes offrent relativement peu d'intérêt, si l'on excepte le fait qu'ils forment un sous-groupe auquel correspondra toujours une absence de données. Par définition, l'appariement juste ne se prête à aucune mesure corrective qui consisterait à **ajouter** les données manquantes. L'inverse, le non-appariement erroné peut faire l'objet de corrections: collecte de meilleures données, transformation des méthodes d'appariement, amélioration de l'appariement par le biais du couplage synthétique.

Si l'on déclare qu'il n'y a pas eu d'appariement et que l'on apprend par la suite que cette décision était erronée, on devrait se poser les questions suivantes: le non-appariement est-il erroné parce qu'il est impossible de trouver la déclaration d'impôt correspondante même si elle existe? le non-appariement est-il erroné parce que la déclaration d'impôt n'existe pas en dépit du fait qu'elle aurait dû être produite?

Les données sur le revenu associées aux non-appariements se présentent en gros comme suit. Des 31,211 **non-appariements justes**, 19,939 (63.9%) n'avaient aucun revenu soumis à l'impôt et 11,272 (36.1%) avaient un certain revenu, mais la personne qui l'avait touché n'était pas assujettie à l'impôt. Si l'on considère ces 11,272 non-appariements, on observe que 96.5% (10,878 dossiers) des répondants ont rapporté un revenu inférieur à \$2,500 et que 5,929 d'entre eux ont déclaré un revenu inférieur à \$1,000.

Les **non-appariements erronés** ne comprennent aucun dossier au bas de l'échelle des revenus. Le tableau 19 présente à cet égard une répartition par tranche de revenu et statut d'appariement dans laquelle l'univers potentiel est composé de l'ensemble des dossiers appariés et non appariés. Les dossiers non appariés sont ventilés en deux catégories, justes et erronés.

En ce qui concerne les **non-appariements justes**, on observe qu'un manque de 19.0% dans le nombre des dossiers entraîne un déficit de 4.0%

as high as 63.3% for the \$1 to \$500 income class and the income shortfall within this class is 58.6%. Up to \$1,500, all classes are reduced to less than one half. In other words, the matched set, due to Revenue Canada coverage limitations, only accounts for every second person at the low end of the distribution, although the aggregate income effect at 4.0% is hardly noticeable.

False non-matches are more difficult to assess since the non-match decision is not necessarily caused by absence of the record in the tax universe. Our inability to link these records affects 6.9% of all potential income recipients, but it produces an income shortfall of 9.1%. The effect on individual income classes is fairly uniform. Membership is affected by less than 10.0% in most cases, and the income effect usually corresponds closely to the membership effect for each class interval.

The impact of true non-matches on statistical output is more damaging than that of false non-matches, since the resulting deficiency cannot be remedied. The high concentration of true non-matches in the lower income classes distorts relative income shares disproportionately. Non-matching due to non-filing or unidentifiable data, as manifested by false non-matches, is distributed more uniformly between income classes. Consequently, adjustments can be made on the basis of the known distribution by proportionately adjusting the income series.

Out of 4,084 false non-matches, 3,637 (89.1%) have wages and salaries as their major source of income. Since income tax was withheld at the source, it is conceivable that the recipient did not file a tax return because a tax liability was not perceived.

It must be stressed, however, that additional tax liabilities due to secondary income sources did arise in many instances. There are 3,637 non-matched wage earners who also have income from non-farm self-employment in 115 instances, from farming in 43 cases, from old-age security in 191 cases, and from pensions on 54 occasions. Investment income is present for 470 of these major wage earners and other income subject to taxation was reported in 69 instances. The 942 cells showing secondary income sources are not mutually exclusive and do indicate how many of the 3,637 wage

au niveau du revenu de l'univers; l'effet atteint un sommet de 63.3% dans la classe de revenu de \$1 à \$500, le déficit correspondant étant de 58.6%. Jusqu'à \$1,500, toutes les classes sont réduites à moins de la moitié. En d'autres termes, bien que l'ensemble des dossiers appariés ne couvre qu'une personne sur deux au bas de l'échelle des revenus (la situation est imputable aux limites du taux de couverture de Revenu Canada), le revenu global n'en souffre pratiquement pas (4.0%).

Il est plus difficile d'évaluer les **non-appariements erronés**, car la décision sur laquelle ils reposent n'est pas nécessairement fondée sur l'absence des dossiers de l'univers de l'impôt. L'impossibilité de lier ces dossiers touche 6.9% des personnes observées, mais entraîne une différence en moins de 9.1%. L'effet sur chacune des classes de revenu est relativement uniforme. Dans la plupart des cas, la différence en moins en ce qui concerne les dossiers est inférieure à 10.0%; les effets observés sur le revenu suivent d'assez près.

L'incidence des non-appariements justes sur les résultats statistiques est plus préjudiciable que celle des non-appariements erronés, car la différence en moins qui en résulte ne peut pas être corrigée. La forte concentration des non-appariements justes dans les classes de revenu inférieures déforme inégalement les parts relatives occupées par le revenu. Les non-appariements attribuables à la non-production d'une déclaration ou à l'impossibilité d'identifier des données (non-appariements erronés) sont répartis plus également. On peut donc ajuster proportionnellement les séries sur le revenu en s'appuyant sur les répartitions connues.

Des 4,084 non-appariements erronés, 3,637 (89.1%) correspondaient à des dossiers dans lesquels les rémunérations constituaient la principale source de revenu. Comme l'impôt sur le revenu a été prélevé à la source, il est possible que certains salariés n'aient pas produit leur déclaration parce qu'ils ne se sentaient plus assujettis à l'impôt.

Toutefois, il convient de souligner que, dans bon nombre de cas, les répondants tiraient des revenus secondaires d'autres sources. Ainsi, des 3,637 salariés non appariés, 115 tiraient également un revenu d'un travail autonome non agricole, 43, de l'agriculture, 191, de prestations de sécurité de la vieillesse et 54, de pensions. De plus 470 de ces salariés avaient eu des revenus de placements et 69, d'autres revenus soumis à l'impôt. Les 942 cas où il y a eu revenus secondaires ne s'excluent pas mutuellement; il n'est donc pas possible de savoir combien de salariés sur les 3,637 avaient des revenus secondaires. Au mieux, ce chiffre constitue une limite

earners have secondary sources. At best, it is an upper limit. In other words, not more than 25% of these wage earners would have incurred additional tax liabilities after withholding taxes.

There are also 447 false non-matches with major income sources not subject to withholding tax; e.g., 235 have non-farm self-employment as a major source, 111 derive the largest income share from farming, 61 from pensions and old-age security, and the remaining 40 records have no unique major source; i.e., two or more income sources are of the same magnitude, or the income components fall into "miscellaneous" categories.

The size classes of "income subject to taxation" for false non-matches with wages as a major source are under \$2,000 in 269 cases, fall between \$2,001 and \$5,000 in 1,470 cases, and are greater than \$5,000 in 1,898 instances.

The income class membership of major source records not subject to withholding taxes falls into the range \$1,000 to \$2,000 for 30 records, into the range \$2,001 to \$5,000 for 242 records, and above \$5,000 for 175 records.

The foregoing discussion of non-matches concludes the technical part of this report. A few points are worth repeating. These points concern data quality in the context of record linkage. It also seems appropriate to reiterate notions surrounding record linkage as a useful tool in the statistician's workshop. The following section will be devoted to these subjects.

Postscript

The desirability of employing up-to-date technology for the improvement and production of statistical output seems to require little justification. "Up till now, technological advance has helped to increase productivity by providing the same output at reduced per-unit costs, or even more output at reduced cost. Often, new technology helped to improve timeliness. Data quality was controlled by way of sampling design, collection procedures, consistency checks, as well as edits and imputations. This approach is still valid and cost-effective for large aggregates where random shocks will cancel out; it is also acceptable, with reservations, for most cross-section and time series data. However, where technology has brought the advent of machine-readable micro data sets, consisting of individual records, and where present-day technology has provided the means of producing longitudinally linked records, conventional standards of data quality should be challenged.

maximale. En d'autres termes, il n'est pas possible que plus de 25% de ces salariés aient touché des revenus additionnels en sus de ceux qui ont fait l'objet d'une retenue à la source.

On dénombre également 447 non-appariements erronés dans lesquels la principale source de revenu n'a pas fait l'objet d'une retenue d'impôt à la source: 235 des répondants ont tiré leur principale source de revenu d'un travail autonome non agricole, 111, de l'agriculture et 61, de pensions et de prestations de sécurité de la vieillesse; dans les 40 dossiers restants, il n'y avait pas de principale source de revenu unique (sources de revenu sensiblement égales, éléments du revenu tombant dans la catégorie "divers").

Les classes de taille du "revenu soumis à l'impôt" des non-appariements erronés dans lesquels la rémunération constitue la principale source de revenu sont inférieures à \$2,000 dans 269 cas, se situent entre \$2,001 et \$5,000 dans 1,470 cas se sont supérieures à \$5,000 dans 1,898 cas.

Le nombre des dossiers dans lesquels la principale source de revenu n'a pas fait l'objet de retenues à la source se situe à 30 dans la classe \$1,001-\$2,000, 242 dans la classe \$2,001-\$5,000 et 175 dans la classe de \$5,000 et plus.

Cet examen des non-appariements met fin à la partie technique de cette étude. Certains points méritent d'être répétés. Ils concernent surtout la qualité des données dans le contexte du couplage des dossiers. Il semble également intéressant de rappeler quelques faits au sujet des avantages statistiques du couplage des dossiers. C'est ce à quoi sera consacrée la section qui suit.

Post-scriptum

La nécessité d'utiliser des techniques modernes pour améliorer et produire des données statistiques n'a pratiquement pas à être justifiée. Jusqu'à ce jour, les progrès techniques nous ont permis d'accroître notre productivité en produisant les mêmes chiffres à un coût réduit voire même en produisant plus à un coût réduit. Les techniques modernes nous ont souvent aidées à améliorer l'actualité des données. La qualité des données a pu être contrôlée par le truchement du plan de sondage, des méthodes de collecte, des vérifications de compatibilité, des contrôles des imputations. Cette approche demeure toujours rentable dans les grands agrégats où les erreurs aléatoires s'annulent; elle est également acceptable, sous certaines réserves, dans la plupart des études transversales et des séries chronologiques. Toutefois, dans les secteurs où la technique a permis l'utilisation de micro-données exploitables par une machine et la production de dossiers couplés longitudinalement, les normes habituelles de qualité des données devraient être remises en question.

Present-day trends may help to place greater emphasis on data quality. In everyday life, such notions as "small is better", "quality of life", and a shift from consumption to conservation have influenced people's actions. Similarly, in the public domain, the need to conserve, to recycle, and to stress quality of service have occupied centre stage. Consequently, record linkage in conjunction with the exploitation of administrative data, such as tax records, conforms to accepted notions of conservation and recycling; the expected or resulting improvement in data quality will depend on the particular administrative files to be used.

To enhance the information value of diverse data sources without adding to response burden, record linkage can be an indispensable tool. Having accepted this premise, data quality of relatively small source files can be improved in key areas, and such improved quality would aid the linkage process as well as enhance the quality and information value of the enlarged linked file.

Record linkage, as described in the body of this report, can be employed as a **substitute** for conventional data collection, or it may **complement** conventionally assembled data files. It was shown in the post-match analysis that substitution of tax data for Census data would have left unacceptable gaps in our statistical knowledge, although as complementary data, the information obtained from record linkage was most useful. Looking at future applications as data substitution or complementation, a number of scenarios can be anticipated.

One possible scenario envisages greater use of administrative records for intercensal years. Record linkage of such administrative data with appropriate Census records would establish bench marks for the base year, and the relationships established in this fashion could be used to make adjustments to data derived solely from administrative sources for intercensal years.

The Census may not always lend itself to record linkage applications, because it is an instrument designed to have the widest possible application. Where a specialized instrument is needed, appropriately designed surveys could form the foundation for a linked data base. Such a linked data base could then be augmented with administrative data for a number of years, thereby requiring a survey less frequently than would be the case under conventional operating procedures.

Les tendances actuelles font qu'on insiste davantage sur la qualité des données. Dans la vie de tous les jours, la recherche d'une société à la mesure de l'individu, l'accent mis sur la qualité de la vie et le passage de la consommation à la conservation ont influencé nos actions. Parallèlement, dans le domaine public, le besoin de conserver, de recycler et d'accroître la qualité des services viennent au premier rang. En ce sens, le couplage des dossiers et l'exploitation des données administratives telles que les déclarations d'impôt s'intègrent bien au concept de la conservation et du recyclage. L'optimisation de la qualité des données qui en résulteront dépend néanmoins des fichiers administratifs utilisés.

Le couplage des dossiers constitue un excellent moyen d'accroître la valeur informative des diverses sources de données sans pour autant alourdir le fardeau des répondants. Cette prémisses étant acceptée, il est possible d'améliorer la qualité de certaines données clés de fichiers de base relativement petits; cela facilitera le processus de couplage et accroîtra la qualité et la valeur informative du fichier ainsi obtenu.

Le couplage des dossiers décrit dans les pages qui précèdent peut donc **remplacer** les méthodes habituelles de collecte des données ou **compléter** les fichiers constitués de la manière habituelle. Nous avons vu dans l'analyse post-appariement que la substitution des données fiscales à celles du recensement aurait laissé des lacunes inacceptables; en revanche, les renseignements obtenus à la suite du couplage ont une utilité indéniable. Si l'on songe aux possibilités d'application du couplage des données, plusieurs scénarios s'offrent à nous.

On pourrait d'une part songer à mettre davantage à profit les dossiers administratifs pendant les périodes intercensitaires. Le couplage de ces données et des dossiers du recensement nous permettrait d'établir des points repères pour l'année de référence; les liens ainsi établis pourraient à leur tour servir à ajuster les données intercensitaires tirées uniquement de dossiers administratifs.

Le recensement ne se prête pas toujours à l'appariement des dossiers, car il a été conçu pour avoir le plus vaste champ d'application possible. Si l'on a besoin d'instruments spécialisés, on pourra alors constituer la base de données nécessaires au couplage par le biais d'enquêtes spécialement conçues. Cette base de données couplées pourra ensuite être augmentée au moyen de dossiers administratifs pendant un certain nombre d'années, ce qui réduira la fréquence des enquêtes.

All linkage activities must be restricted to samples of the population. The cost of large-scale linkages is still prohibitive. A relatively small sample, made up of high-quality data, and processed under stringent quality control measures is most promising, given the present state of the arts, and the desire to keep expenditure low.

Toute activité de couplage doit être limitée à des échantillons de la population. Le coût des couplages à grande échelle demeure en effet prohibitif. Compte tenu de l'état actuel de nos connaissances et de nos ressources financières, on cherchera donc à s'appuyer sur les petits échantillons constitués de données de grande qualité et dont l'exploitation devrait faire l'objet de mesures de contrôle qualitatif très sévères.

Footnotes

- (1) Benjamin Okner, "Constructing a New Data Base from Existing Micro-data Sets: the 1966 Merge File", in **Annals of Economics and Social Measurement**, July 1972.
- (2) Horst Alter, "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970", in **Annals of Economic and Social Measurement**, April 1974.
- (3) United States Department of Commerce, Bureau of the Census, "Some Preliminary Results from the 1973 CPS-IRS-SSA Exact Match Study" - mimeo, Sept. 1975. This selection of papers contains a vast number of bibliographical references to other United States Government publications resulting from these linkage activities.

A valuable summary, but to-date unpublished, can be expected under the authorship of Beth Kilss and F. Scheuren, Office of Research and Statistics, Social Security Administration. They presented a preliminary version of "The 1973 CPS-IRS-SSA Exact Match Study - Past, Present and Future" at the NBER Workshop on the Policy Uses of Social Security Research Files, March 15-17, 1978.
- (4) J.M. Kennedy, H.B. Newcombe, E.A. Okazaki, and M.E. Smith, "Computer Methods for Family Linkage of Vital and Health Records", Atomic Energy of Canada Limited, Chalk River, Ontario, April 1965. There are various articles in professional journals by Newcombe, Smith and Kennedy under single or joint authorship on various aspects of record linkage. Please refer to the Bibliography.
- (5) I.P. Fellegi and A.B. Sunter, "A Theory for Record Linkage" in the **Journal of the American Statistical Association**, Vol. 64, 1969, pp. 1183-1210.
- (6) While observed differences in income are interpreted as reporting errors, these differences may result from flaws in data capture and processing, or they may be the result of edits and imputations on Census forms. It should also be noted that tax data used do not reflect late taxfilers, re-assessment, or supplementary corrections filed after the cut-off date of the file creation.

Notes

- (1) Benjamin Okner, "Constructing a New Data Base from Existing Micro-data Sets: the 1966 Merge File", **Annals of Economics and Social Measurement**, juillet 1972.
- (2) Horst Alter, "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970", **Annals of Economic and Social Measurement**, avril 1974.
- (3) Ministère du Commerce des États-Unis, Bureau du recensement, "Some Preliminary Results from the 1973 CPS-IRS-SSA Exact Match Study" - ronéo, septembre 1975. Ces articles contiennent une abondante bibliographie consacrée à d'autres publications du gouvernement américain produites à la suite de ces travaux de couplage.

Les résultats devraient être présentés de façon sommaire par Beth Kilss et F. Scheuren, Office of Research and Statistics, Social Security Administration. Ils ont présenté une version préliminaire de "The 1973 CPS-IRS-SSA Exact Match Study - Past, Present and Future" lors du colloque de la NBER sur les conséquences politiques de l'utilisation des fichiers de recherche de la Sécurité sociale (15 au 17 mars 1978).
- (4) J.M. Kennedy, H.B. Newcombe, E.A. Okazaki et M.E. Smith, "Computer Methods for Family Linkage of Vital and Health Records", Énergie atomique du Canada Limitée, Chalk River, Ontario, avril 1965. Newcombe, Smith et Kennedy ont publié un certain nombre d'articles sur les aspects du couplage des dossiers dans diverses revues spécialisées.
- (5) I.P. Fellegi et A.B. Sunter, "A Theory for Record Linkage" **Journal of the American Statistical Association**, vol. 64, 1969, pp. 1183 à 1210.
- (6) Bien que ces écarts puissent être assimilés à des erreurs de déclaration, ils peuvent également provenir d'erreurs de saisie et d'exploitation des données ou résulter des procédures de contrôle et d'imputation. Il faut aussi noter que les données fiscales ne tiennent pas compte des déclarants retardataires, des réévaluations et des corrections additionnelles introduites après la date limite de création du fichier.

- (7) A detailed discussion of the "Utility derived from longitudinal data" is contained in an unpublished staff paper, prepared in July 1980 in the Consumer Income and Expenditure Division of Statistics Canada. Its contents will be incorporated in a future publication of longitudinal data and analysis.
- (8) The use of apartment number was intended, but processing problems prevented our using it.
- (9) Actually, the so-called REDID was used which consists of the first five characters of the surname with some modifications for blanks and apostrophies and a uniform treatment of Mc... and Mac... .
- (10) All figures are rounded to the nearest thousand.
- (11) See introductory section - The Matching of Tax and Census Records.
- (12) These very stringent conditions apply to the first round. In the second round, only surname had to agree and year of birth and month of birth could deviate within certain limits.
- (13) While we succeeded in isolating apartment numbers, we ultimately failed in processing them properly and had to do without them. Thus, apartment numbers will not be discussed any further. The pilot project, however, showed that apartment numbers are a valuable data item in reaching matching decisions.
- (14) Expected total income was estimated as follows:

	\$'000,000
Census - Recensement	246.131
RCR - RCI	239.707

- (7) Une discussion détaillée de l'utilité des données longitudinales figure dans un document non publié rédigé en juillet 1980 par la Division du revenu et des dépenses des consommateurs de Statistique Canada. Son contenu sera incorporé à une publication ultérieure consacrée à l'analyse et aux données longitudinales.
- (8) Certains problèmes d'exploitation nous ont empêchés d'utiliser le numéro d'appartement.
- (9) En fait, le variable REDID utilise les cinq premiers caractères du nom de famille, compte tenu de certaines modifications destinées à tenir compte des blancs, des apostrophes et du traitement des Mc et Mac.
- (10) Les chiffres sont arrondis au millier près.
- (11) Voir, dans les remarques liminaires, la section intitulée Appariement des dossiers de l'impôt et du recensement.
- (12) Ces conditions très sévères ne s'appliquaient qu'à la première série d'interrogations. Dans la deuxième série, seul le nom de famille devait coïncider; l'année et le mois de naissance pouvaient s'écarter quelque peu.
- (13) Bien que nous ayons réussi à isoler les numéros d'appartement, nous ne sommes pas parvenus à les exploiter correctement et nous avons dû nous en passer. La question des numéros d'appartement ne sera donc pas examinée plus à fond. Le projet pilote montrait néanmoins que le numéro d'appartement constituait une donnée fort utile pour en arriver à une décision quant à l'appariement.
- (14) Le revenu total prévu a été estimé comme suit:

Conceptually compatible total income for all true matches, but some of it reported in one source only - Total conceptuellement compatible pour l'ensemble des appariements justes; certains éléments du revenu n'ont toute fois été déclarés que dans une seule source

	\$'000,000	
Census (non-matches) -	36.163	(36.163 x 0.0261 =
Recensement (non-	<u>.944</u>	
appariements)		
Est. RC-T non-match		estimated overreport
component - Non-		ing -) surdéclaration
appariements RC-I est.	35.219	estimative
RC-T matches -		
Appariements RC-I	<u>239.707</u>	
	<u>274.926</u>	

(15) Roger A. Herriot and Emmelt F. Spiers, "Some Preliminary Results from the 1973 CPS-IRS-SSA Exact Match Study", United States Department of Commerce, Bureau of the Census - mimeo. This is one of the papers delivered at the 1975 annual meeting of the American Statistical Association and was to appear in the 1975 Proceedings of the Social Statistics Section.

(15) Roger A. Herriot et Emmelt F. Spiers, "Some Preliminary Results from the 1973 CPS-IRS-SSA Exact Match Study", ministère du Commerce des États-Unis, Bureau du recensement, ronéo. Il s'agit d'un des exposés présentés à l'occasion de la réunion annuelle de 1975 de la American Statistical Association; il figure dans le procès-verbal de la Section des statistiques sociales.

From Table 5 of that paper, the following results have been reworked and can be compared with Table 18 of this paper. "Census Overreported" should read CPS overreported, and "Census Underreported" should read CPS underreported in the United States context, where CPS is the Current Population Survey.

Les chiffres qui suivent sont tirés du tableau 5 de ce document; ils ont été reformulés de façon à pouvoir être comparés à ceux du tableau 18 de notre étude. Les titres "surdéclaration - recensement" et "sous-déclaration - recensement" du tableau 18 correspondent respectivement à la surdéclaration et à la sous-déclaration dans l'enquête américaine (Current Population Survey).

More than one class below main diagonal, 2,081 or 5.3%.

Plus de une classe sous la diagonale principale, 2,081 ou 5.3%.

One class below main diagonal, 3,145 or 8.0%.

Une classe sous la diagonale principale, 3,145 ou 8.0%.

On main diagonal, 25,618 or 65.2%.

Sur la diagonale principale, 25,618 ou 65.2%.

One class above main diagonal, 4,784 or 12.2%.

Une classe au-dessus de la diagonale principale, 4,784 ou 12.2%.

More than one class above main diagonal, 3,645 or 9.3%.

Plus de une classe au-dessus de la diagonale principale, 3,645 ou 9.3%.

TABLE 1. Census Income Recipients, by Match Status and by Major Source of Income for Income Base Year, 1970

TABLEAU 1. Personnes ayant déclaré un revenu au recensement, selon le statut d'appariement et la principale source de revenu, 1970

Major source of income Principale source de revenu	Match Appariement		Non-match Non-appariement		Total	
	number	per cent	number	per cent	number	per cent
	nombre	pourcentage	nombre	pourcentage	nombre	pourcentage
Wages and salaries - Rémunérations	35,276	82.6	7,383	48.1	42,659	73.5
Income from self-employment - Revenu d'un travail autonome	2,527	5.9	719	4.7	3,246	5.6
Multiple-earned income(1) - Revenu gagné tiré de plusieurs sources(1)	172	0.4	83	0.5	255	0.4
All earned income - Total, revenu gagné	37,975	88.9	8,185	53.5	46,160	79.5
Non-earned income(2) - Revenu non gagné(2)	4,738	11.1	7,172	46.7	11,910	20.5
All major sources - Total, principales sources	42,713	100.0	15,357	100.0	58,070	100.0

(1) Two or more sources are of equal size and occupy top rank.

1) Deux ou plusieurs sources de taille égale venant au premier rang.

2) Summarized, but defined for the following sources: Retirement Income (combined old-age security and pension income), Miscellaneous (includes investment income, rental income and "Other" income).

(2) Comprend les sources suivantes: retraite (prestations de sécurité de la vieillesse et pensions), divers (revenus de placements, revenus locatifs et "autres revenus").

TABLE 2. Adult Census Population by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971

TABLEAU 2. Population adulte du recensement par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971

Age group Groupe d'âge	Male Masculin		Female Féminin		Male and female Masculin et féminin		
	Married	Not married	Married	Not married	Total	Married	Not married
	Marié	Non marié	Mariée	Non mariée		Marié	Non marié
15-20 years - ans	76,230	1,186,561	181,289	1,048,157	2,492,237	257,519	2,234,718
21-64 years - ans	4,250,960	1,236,279	4,330,496	1,133,003	10,950,738	8,581,456	2,369,282
65 years and over - ans et plus	561,570	220,290	377,055	585,515	1,744,430	938,625	805,805
All ages - Tout âges	4,888,760	2,643,130	4,888,840	2,766,675	15,187,405	9,777,600	5,409,805
Percentage distribution between age groups - Répartition en pourcentage par groupe d'âge:							
15-20 years - ans	1.6	44.9	3.7	37.9	16.4	2.6	41.3
21-64 years - ans	87.0	46.8	88.6	41.0	72.1	87.8	43.8
65 years and over - ans et plus	11.5	8.3	7.7	21.2	11.5	9.6	14.9
All ages - Tout âges	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Percentage distribution within age groups - Répartition en pourcentage par état matrimonial et sexe:							
15-20 years - ans	3.1	47.6	3.3	42.1	100.0	10.3	89.7
21-64 years - ans	38.8	11.3	39.5	10.3	100.0	78.4	21.6
65 years and over - ans et plus	32.2	12.6	21.6	33.6	100.0	53.8	46.2
All ages - Tout âges	32.2	17.4	32.2	18.2	100.0	64.4	35.6

Note: Age and marital status as of June 1, 1971.

Nota: Âge et état matrimonial au 1^{er} juin 1971.

Source: Census of Population, 1971 (Statistics Canada); with interpolation for 20-year age group from published data.

Source: Recensement de la population, 1971 (Statistique Canada); les données du groupe d'âge de 20 ans ont été interpolées à partir de données publiées.

TABLE 3. Sample of Adult Census Population by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971

TABLEAU 3. Échantillon de la population adulte du recensement par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971

Age group Groupe d'âge	Male Masculin		Female Féminin		Male and female Masculin et féminin		
	Married Marié	Not married Non marié	Married Mariée	Not married Non mariée	Total	Married Marié	Not married Non marié
15-20 years - ans	232	6,712	850	6,085	13,879	1,082	12,797
21-64 years - ans	21,305	6,646	21,667	6,435	56,053	42,972	13,081
65 years and over - ans et plus	2,733	1,382	1,754	3,380	9,249	4,487	4,762
All ages - Tout âges	24,270	14,740	24,271	15,900	79,181	48,541	30,640
Percentage distribution between age groups - Répartition en pourcentage par groupe d'âge:							
15-20 years - ans	1.0	45.5	3.5	38.3	17.5	2.2	41.8
21-64 years - ans	87.8	45.1	89.3	40.5	70.8	88.5	42.7
65 years and over - ans et plus	11.3	9.4	7.2	21.3	11.7	9.2	15.5
All ages - Tout âges	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Percentage distribution within age groups - Répartition en pourcentage par état matrimo- nial et sexe:							
15-20 years - ans	1.7	48.4	6.1	43.8	100.0	7.8	92.2
21-64 years - ans	38.0	11.9	38.6	11.5	100.0	76.7	23.3
65 years and over - ans et plus	29.6	14.9	19.0	36.5	100.0	48.5	51.5
All ages - Tout âges	30.6	18.6	30.6	20.1	100.0	61.3	38.7

Note: Age and marital status as of June 1, 1971.

Nota: Âge et état matrimonial au 1^{er} juin 1971.

TABLE 4. Matched Records(1) by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971

TABLEAU 4. Dossiers appariés(1) par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971

Age group Groupe d'âge	Male Masculin		Female Féminin		Male and female Masculin et féminin		
	Married Marié	Not married Non marié	Married Mariée	Not married Non mariée	Total	Married Marié	Not married Non marié
15-20 years - ans	188	2,359	362	1,626	4,535	550	3,985
21-64 years - ans	18,855	4,568	8,084	3,945	35,452	26,939	8,513
65 years and over - ans et plus	1,827	547	391	1,134	3,899	2,218	1,681
All ages - Tout âges	20,870	7,474	8,837	6,705	43,886	29,707	14,179
Percentage distribution between age groups - Répartition en pourcentage par groupe d'âge:							
15-20 years - ans	9.9	31.6	4.1	24.3	10.3	1.9	28.1
21-64 years - ans	90.3	61.6	91.5	58.8	80.8	90.7	60.0
65 years and over - ans et plus	8.8	7.3	4.4	16.9	8.9	7.5	11.9
All ages - Tout âges	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Percentage distribution within age groups - Répartition en pourcentage par état matrimo- nial et sexe:							
15-20 years - ans	4.1	52.0	8.0	35.9	100.0	12.1	87.9
21-64 years - ans	51.2	17.9	22.8	11.1	100.0	70.0	24.0
65 years and over - ans et plus	44.9	14.0	10.6	29.1	100.0	56.9	43.1
All ages - Tout âges	47.6	17.0	20.1	15.3	100.0	67.7	32.3

(1) True matches only, since these data were compiled after editing out and converting false matches to non-matches.

(1) Appariements justes seulement; la compilation des données s'est en effet faite après la vérification et la conversion des appariements erronés en non-appariements.

Note: Age and marital status as of June 1, 1971.

Nota: Âge et état matrimonial au 1^{er} juin 1971.

TABLE 5. True Matches with Income Reported in One Source(1) Only, by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971

TABLEAU 5. Appariements justes, revenu déclaré dans une seule source(1) par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971

Age group Groupe d'âge	Male Masculin		Female Féminin		Male and female Masculin et féminin		
	Married Marié	Not married Non marié	Married Mariée	Not married Non mariée	Total	Married Marié	Not married Non marié
15-20 years - ans	1	226	23	126	376	24	352
21-64 years - ans	117	125	455	89	786	572	214
65 years and over - ans et plus	1	3	1	4	9	2	7
All ages - Tout âges	119	354	479	219	1,171	598	573
Percentage distribution between age groups - Répartition en pourcentage par groupe d'âge:							
15-20 years - ans	0.2	63.8	4.8	57.5	32.1	4.0	61.4
21-64 years - ans	98.3	35.3	95.0	40.6	67.1	95.7	37.3
65 years and over - ans et plus	0.8	0.8	0.2	1.8	0.8	0.3	1.2
All ages - Tout âges	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Percentage distribution within age groups - Répartition en pourcentage par état matrimo- nial et sexe:							
15-20 years - ans	6.3	60.1	6.1	33.5	100.0	6.4	93.6
21-64 years - ans	14.4	15.9	57.9	11.3	100.0	72.8	27.2
65 years and over - ans et plus	11.1	33.3	11.1	44.4	100.0	22.2	77.8
All ages - Tout âges	10.2	30.2	40.9	18.7	100.0	51.1	48.9

(1) Census or Revenue Canada - Taxation (RC-T), but most of these records (1158/1171) have reported income exclusively to RC-T.

(1) Recensement ou Revenu Canada - Impôt (RC-I); dans la plupart des cas (1158/1171), toutefois, le revenu n'avait été déclaré qu'à RC-I.

Note: Age and marital status as of June 1, 1971.

Nota: Âge et état matrimonial au 1^{er} juin 1971.

TABLE 6. All Non-matches by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971

TABLEAU 6. Ensemble des non-appariements par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971

Age group Groupe d'âge	Male Masculin		Female Féminin		Male and female Masculin et féminin		
	Married Marié	Not married Non marié	Married Mariée	Not married Non mariée	Total	Married Marié	Not married Non marié
15-20 years - ans	44	4,353	488	4,459	9,344	532	8,812
21-64 years - ans	2,450	2,078	13,583	2,490	20,601	16,033	4,568
65 years and over - ans et plus	906	835	1,363	2,246	5,350	2,269	3,081
All ages - Tout âges	3,400	7,266	15,434	9,195	35,295	18,834	16,461
Percentage distribution between age groups - Répartition en pourcentage par groupe d'âge:							
15-20 years - ans	1.3	59.9	3.2	48.5	26.5	2.8	53.5
21-64 years - ans	71.1	28.6	88.0	27.1	58.5	85.1	22.8
65 years and over - ans et plus	27.6	11.5	8.8	24.4	15.2	12.0	18.7
All ages - Tout âges	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Percentage distribution within age groups - Répartition en pourcentage par état matrimo- nial et sexe:							
15-20 years - ans	1.5	46.6	5.2	47.7	100.0	5.7	94.3
21-64 years - ans	11.9	10.1	65.9	12.1	100.0	77.8	22.2
65 years and over - ans et plus	10.9	15.6	25.5	42.0	100.0	42.4	57.6
All ages - Tout âges	9.6	20.6	43.7	26.1	100.0	53.4	46.6

Note: Age and marital status as of June 1, 1971.

Nota: Âge et état matrimonial au 1^{er} juin 1971.

TABLE 7. True Non-matches by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971

TABLÉAU 7. Non-appariements justes par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971

Age group Groupe d'âge	Male Masculin		Female Féminin		Male and female Masculin et féminin		
	Married Marié	Not married Non marié	Married Mariée	Not married Non mariée	Total	Married Marié	Not married Non marié
15-20 years - ans	25	4,166	427	4,363	8,981	452	8,529
21-64 years - ans	990	1,162	12,980	2,022	17,154	17,970	3,184
65 years and over - ans et plus	823	775	1,336	2,142	5,076	2,159	2,917
All ages - Tout âges	1,838	6,103	14,743	8,527	31,211	16,581	14,630
Percentage distribution between age groups - Répartition en pourcentage par groupe d'âge:							
15-20 years - ans	1.4	68.3	2.9	51.2	28.8	2.7	58.3
21-64 years - ans	53.9	19.0	88.0	23.7	55.0	84.3	21.8
65 years and over - ans et plus	44.8	12.7	9.1	25.1	16.3	13.0	19.9
All ages - Tout âges	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Percentage distribution within age groups - Répartition en pourcentage par état matrimonial et sexe:							
15-20 years - ans	0.3	46.4	4.8	48.6	100.0	5.0	95.0
21-64 years - ans	5.8	6.8	75.7	11.8	100.0	81.4	18.6
65 years and over - ans et plus	16.2	15.3	26.3	42.2	100.0	42.5	57.5
All ages - Tout âges	5.9	19.6	47.2	27.3	100.0	53.1	46.9

Note: Age and marital status as of June 1, 1971.

Nota: Âge et état matrimonial au 1^{er} juin 1971.

TABLE 8. False Non-matches by Basic Age Group, by Sex and Broad Marital Status with Percentage Distributions, 1971

TABLÉAU 8. Non-appariements erronés par grands groupes d'âge et selon le sexe et l'état matrimonial, répartitions en pourcentage, 1971

Age group Groupe d'âge	Male Masculin		Female Féminin		Male and female Masculin et féminin		
	Married Marié	Not married Non marié	Married Mariée	Not married Non mariée	Total	Married Marié	Not married Non marié
15-20 years - ans	19	187	61	96	363	80	283
21-64 years - ans	1,460	916	603	468	3,447	2,063	1,384
65 years and over - ans et plus	83	60	27	104	274	110	164
All ages - Tout âges	1,562	1,163	691	668	4,084	2,253	1,831
Percentage distribution between age groups - Répartition en pourcentage par groupe d'âge:							
15-20 years - ans	1.2	16.1	8.8	14.4	8.9	3.6	15.5
21-64 years - ans	93.5	78.8	87.3	70.1	84.4	91.6	75.6
65 years and over - ans et plus	5.3	5.2	3.9	15.6	6.7	4.9	9.0
All ages - Tout âges	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Percentage distribution within age groups - Répartition en pourcentage par état matrimonial et sexe:							
15-20 years - ans	5.2	51.5	16.8	26.4	100.0	22.0	78.0
21-64 years - ans	42.4	26.6	17.5	13.6	100.0	59.8	40.2
65 years and over - ans et plus	30.3	21.9	9.9	32.2	100.0	40.1	59.9
All ages - Tout âges	38.2	28.5	16.9	16.4	100.0	55.2	44.8

Note: Age and marital status as of June 1, 1971.

Nota: Âge et état matrimonial au 1^{er} juin 1971.

TABLE 9. Reporting Pattern of Income Components for True Matches, by Source Presence with Consistency Rate for Income Base Year, 1970

TABLEAU 9. Modèle de déclaration des éléments du revenu dans les appariements justes, selon la présence de la source et taux de cohérence, 1970

Income component	Absent in both sources	Present in RC-T record only	Present in Census record only	Present in both sources	
Élément du revenu	Absent dans les deux sources	RC-I seulement	Recensement seulement	Présent dans les deux sources	
	1	2	3	4	
Wages and salaries - Rémunérations	5,032	2,058	1,316	35,480	
Non-farm self-employment income (net) - Revenu non agricole d'un travail autonome (net)	39,535	1,405	1,214	1,732	
Farm self-employment income (net) - Revenu agricole d'un travail autonome (net)	41,622	432	334	1,498	
Old-age security income - Prestations de sécurité de la vieillesse	39,768	137	233	3,748	
Pension income - Pensions	40,869	816	718	1,483	
Investment income(1) - Placements(1)	23,906	7,756	1,303	10,921	
Other income - Autres revenus	39,216	3,668	691	311	
Income subject to taxation - Revenu soumis à l'impôt	4	1,158	13	42,711	
	Present in at least one source	Consistently present or absent in both sources	Number of cells compared	Consistency rate(2)	Single-it rate(3)
	Présent dans au moins une source	Systématiquement présent ou absent dans les deux sources	Nombre de dossiers comparés	Taux de cohérence(2)	Taux de présence unique(3)
	5	6	7	8	9
per cent - pourcentage					
Wages and salaries - Rémunérations	38,854	40,512	43,886	92.3	7.7
Non-farm self-employment income (net) - Revenu non agricole d'un travail autonome (net)	4,351	41,267	43,886	94.0	6.0
Farm self-employment income (net) - Revenu agricole d'un travail autonome (net)	2,264	43,120	43,886	98.2	1.8
Old-age security income - Prestations de sécurité de la vieillesse	4,118	43,516	43,886	99.2	0.8
Pension income - Pensions	3,017	42,352	43,886	96.5	3.5
Investment income(1) - Placements(1)	19,980	34,827	43,886	79.4	20.6
Other income - Autres revenus	4,670	39,527	43,886	90.1	9.9
Income subject to taxation - Revenu soumis à l'impôt	43,882	42,715	43,886	97.3	2.7

(1) Includes net rental income.

(1) Revenus locatifs nets compris.

(2) The sum of all records with component absent in both sources (Column 1), and records with component present in both sources (Column 4), all divided by the number of cells compared (Column 7).

(2) Somme des dossiers dans lesquels le revenu est absent dans les deux sources (colonne 1) et de ceux dans lesquels il est présent dans les deux sources (colonne 4) divisée par le nombre de chiffres comparés (colonne 7).

(3) The sum of all records with component present exclusively in RC-T, or exclusively in Census, all divided by the number cells compared.

(3) Somme de l'ensemble des dossiers dans lesquels le revenu n'est présent que dans une seule source divisée par le nombre chiffres comparés.

TABLE 10. Reporting Patterns of Income Components, by Province for True Matches with Income Subject to Taxation Present in Both Sources(1) for the Income Base Year, 1970

TABLEAU 10. Régime de déclaration des éléments du revenu, par province, appariements justes, revenu soumis à l'impôt présent dans les deux sources(1), 1970

Province	Number of records	Total income(2) by source		Non-sampling effect (NSE)	
	Nombre de dossiers	Revenu total(2) par source		Effet d'observation (EO)	
		Census	RC-T	AD(3)	AD/RC-T
		Recensement	RC-I	Écart(3)	Écart/RC-I
		thousands of dollars		per cent	
		milliers de dollars		pourcentage	
					Average AD
					Écart moyen

Records with all components reported consistently in both sources(1) - Dossiers dans lesquels l'ensemble des éléments ont été déclarés uniformément dans les deux sources(1):

Newfoundland - Terre-Neuve	1,285	5,301	5,268	33	0.6	26
Prince Edward Island - Île-du-Prince-Édouard	330	1,295	1,228	67	5.5	203
Nova Scotia - Nouvelle-Écosse	2,441	11,060	11,022	38	0.3	16
New Brunswick - Nouveau-Brunswick	1,990	8,211	8,139	72	0.9	36
Québec	4,982	27,166	27,612	446	1.6	90
Ontario	7,804	44,171	43,860	311	0.7	40
Manitoba	1,524	7,661	7,631	30	0.4	20
Saskatchewan	1,244	5,536	5,193	343	6.6	276
Alberta	2,521	13,405	12,919	486	3.8	193
British Columbia(4) - Colombie-Britannique(4)	3,319	19,133	18,444	689	3.7	208
CANADA	27,440	142,939	141,316	1,623	1.1	59

Records with some components reported inconsistently in both sources(1) - Dossiers dont certains éléments sont déclarés sans uniformité dans les deux sources(1):

Newfoundland - Terre-Neuve	577	3,401	3,204	197	6.1	341
Prince Edward Island - Île-du-Prince-Édouard	167	826	735	91	12.4	545
Nova Scotia - Nouvelle-Écosse	1,299	8,228	7,203	1,025	14.2	789
New Brunswick - Nouveau-Brunswick	913	4,791	4,546	245	5.4	268
Québec	2,758	19,626	18,633	993	5.3	360
Ontario	4,575	33,129	32,059	1,070	3.3	234
Manitoba	867	5,216	4,825	391	8.1	451
Saskatchewan	738	3,777	3,077	700	22.7	949
Alberta	1,455	10,842	9,605	1,237	12.9	850
British Columbia(4) - Colombie-Britannique(4)	1,922	13,314	12,549	765	6.1	398
CANADA	15,271	103,150	96,436	6,714	7.0	440

All true matches with total income	Consistency score(5)	Provincial rank order		
		Rang provincial en ordre		
		Descending consistency score	Ascending average AD	
Ensemble des appariements justes ayant un revenu total	Degré de cohérence(5)	Cohérence (ordre décroissant)	Écart moyen (ordre croissant)	
			Consistent records	Inconsistent records
			Dossiers cohérents	Dossiers incohérents

Records with components in both sources(1) - Dossiers dans lesquels les éléments sont présents dans les deux sources(1):

Newfoundland - Terre-Neuve	1,862	69.0	1	3	3
Prince Edward Island - Île-du-Prince-Édouard	497	66.4	3	8	7
Nova Scotia - Nouvelle-Écosse	3,740	65.3	4	1	8
New Brunswick - Nouveau-Brunswick	2,903	68.5	2	4	2
Québec	7,740	64.4	5	6	4
Ontario	12,379	63.0	9	5	1
Manitoba	2,391	63.7	6	2	6
Saskatchewan	1,982	62.8	10	10	10
Alberta	3,976	63.4	7	7	9
British Columbia(4) - Colombie-Britannique(4)	5,241	63.3	8	9	5
CANADA	42,711	64.2

(1) Census and Revenue Canada - Taxation (RC-T).

(1) Recensement et Revenu Canada - Impôt (RC-I).

(2) Made conceptually compatible; i.e., "income subject to taxation".

(2) Rendu conceptuellement compatible (c.-à-d., "revenu soumis à l'impôt").

(3) The difference in total income (see footnote 2 above) as observed in Census and RC-T records for any given individual, aggregated for each geographical unit and stated in absolute terms.

(3) Écart entre le revenu total (voir note 2) déclaré au recensement et à RC-I par un particulier groupé par unités géographiques et exprimé en chiffres absolus.

(4) Includes records for Yukon and Northwest Territories.

(4) Yukon et Territoires du Nord-Ouest compris.

(5) Number of records with consistent components divided by "all true matches" in per cent.

(5) Nombre de dossiers cohérents divisé par l'ensemble des appariements justes et exprimé en pourcentage.

TABLE 11. Omissions and Substitutions of Income Components, by Reliability Category with Aggregate Total Income by Data Source for the Income Base Year, 1970

TABLEAU 11. Omissions et substitutions d'éléments du revenu, par catégorie de fiabilité, et revenu agrégatif total par source de données, 1970

	Reliability category(1) - Catégorie de fiabilité(1)				Total income(2)	
	Number of records		Total income(2)		Revenu total(2)	
	Nombre de dossiers					
	High (A)	Low (B)	Indeterminate (C)	Total	High (A) - Grande (A)	
	Grande (A)	Faible (B)	Indéterminée (C)		Census	RC-T
					Recensement	RC-I
					thousands of dollars	
					milliers de dollars	
Number of components omitted or substituted and data source - Nombre d'éléments omis ou substitués et source de données:						
One omission - Census - Une omission - Recensement	3,842	2,139	2,640	8,621	23,010	23,12
Two or three omissions - Census - Deux ou trois omissions - Recensement	335	549	383	1,267	2,320	2,33
One omission - RC-T - Une omission - RC-I	552	1,113	591	2,256	3,169	3,15
Two or three omissions - RC-T - Deux ou trois omissions - RC-I	6	140	32	178	39	3
One substitution - Census/RC-T - Une substitution - Recensement/RC-I	850	906	401	2,157	4,968	4,97
Two or three substitutions - Census/RC-T - Deux ou trois substitutions - Recensement/RC-I	13	27	8	48	86	8
One or two omissions (Census) and one or two substitutions - Une ou deux omissions (recensement) et une ou deux substitutions	115	321	140	576	734	73
One or two omissions (RC-T) and one or two substitutions - Une ou deux omissions (RC-I) et une ou deux substitutions	13	112	33	158	89	8
Other multiple omissions and substitutions - Autres omissions et substitutions multiples	1	9	-	10	5	
All records with omissions and/or substitutions - Ensemble des dossiers avec omissions et (ou) substitutions	5,727	5,316	4,228	15,271	34,420	34,53
All records without omissions and/or substitutions - Ensemble des dossiers sans omissions et (ou) substitutions	17,244	4,580	5,616	27,440	84,722	84,85
All true matches with income subject to taxation in both sources (Census and RC-T) - Ensemble des appariements justes dont le revenu est soumis à l'impôt dans les deux sources (recensement et RC-I)	22,971	9,896	9,844	42,711	119,142	119,39

See footnote(s) at end of table.
Voir note(s) à la fin du tableau.

BLE 11. Omissions and Substitutions of Income Components, by Reliability Category with Aggregate Total Income by Data Source for the Income Base Year, 1970 - Concluded

BLEAU 11. Omissions et substitutions d'éléments du revenu, par catégorie de fiabilité, et revenu agrégatif total par source de données, 1970 - fin

Reliability category(1) - Catégorie de fiabilité(1)						
Total income(2) - Revenu total(2)						
Low (B)		Indeterminate (C)		Total		
Faible (B)		Indéterminée (C)				
Census	RC-T	Census	RC-T	Census	RC-T	
Recensement	RC-I	Recensement	RC-I	Recensement	RC-I	
thousands of dollars - milliers de dollars						
Number of components omitted or substituted and data source - Nombre d'éléments omis ou substitués et source de données:						
One omission - Census - Une omission - Recensement	10,640	11,677	20,586	21,385	54,236	56,183
One or three omissions - Census - Deux ou trois omissions - Recensement	2,765	3,798	3,715	3,816	8,800	9,948
One omission - RC-T - Une omission - RC-I	9,555	4,784	4,260	4,079	16,984	12,019
One or three omissions - RC-T - Deux ou trois omissions - RC-I	2,228	618	281	282	2,548	939
One substitution - Census/RC-T - Une substitution - Recensement/RC-I	6,523	3,920	3,090	3,098	14,581	11,988
One or three substitutions - Census/RC-T - Deux ou trois substitutions - Recensement/RC-I	292	194	44	47	422	327
One or two omissions (Census) and one or two substitutions - Une ou deux omissions (recensement) et une ou deux substitutions	2,150	1,979	1,226	1,235	4,110	3,952
One or two omissions (RC-T) and one or two substitutions - Une ou deux omissions (RC-I) et une ou deux substitutions	970	549	293	275	1,352	912
Other multiple omissions and substitutions - Autres omissions et substitutions multiples	103	157	-	-	108	162
1 records with omissions and/or substitutions - Ensemble des dossiers avec omissions et (ou) substitutions	35,226	27,676	33,495	34,217	103,141	96,430
1 records without omissions and/or substitutions - Ensemble des dossiers sans omissions et (ou) substitutions	20,519	18,091	37,698	38,368	142,939	141,316
1 true matches with income subject to taxation in both sources (Census and RC-T) - Ensemble des appariements justes dont le revenu est soumis à l'impôt dans les deux sources (recensement et RC-I)	55,745	45,767	71,193	72,585	246,080	237,746

The classification is based on the absolute reporting error as well as on the percentage error with RC-T data as the base. The reporting error has been calculated for conceptually compatible "total income" from Census and RC-T sources. Reliability of reporting is **high** (A), whenever the absolute error is \$200 or less with the percentage error not exceeding 20%. Reliability of reporting is **low** (B), whenever the absolute error exceeds \$200 with the percentage error also exceeding 20%. Reliability of reporting is **indeterminate** (C) for all other records; i.e., whenever a combination of high absolute error and low percentage error occurs (in excess of \$200, but less than 20%), or of low absolute error but high percentage error (not exceeding \$200 but in excess of 20%).

Le classement est fondé sur l'erreur de déclaration absolue et sur le pourcentage d'erreur des données de RC-I. L'erreur de déclaration a été calculée pour un "revenu total" conceptuellement compatible tiré du recensement et de RC-I. La fiabilité est **grande** (A) quand l'erreur absolue est de \$200 ou moins, le pourcentage n'étant pas supérieur à 20%. La fiabilité est **faible** (B) quand l'erreur absolue est supérieure à \$200, le pourcentage d'erreur étant lui aussi supérieur à 20%. La fiabilité est **indéterminée** (C) pour l'ensemble des autres dossiers: erreur absolue élevée et faible pourcentage d'erreur (plus de \$200, mais moins de 20%), faible erreur absolue, mais pourcentage d'erreur élevé (moins de \$200, mais plus de 20%).

Made conceptually compatible; i.e., "income subject to taxation".

Rendu conceptuellement compatible; c.-à-d., "revenu soumis à l'impôt".

TABLE 12. Income Effect of Component Omission by Reliability Category, by Source of Omission, by Incidence Group for Income Base Year, 1970

TABLÉAU 12. Effet sur le revenu de l'omission d'un élément, par catégorie de fiabilité, selon la source de l'omission et le groupe d'incidence, 1970

Incidence group Groupe d'incidence	Reliability category - Catégorie de fiabilité				Total
	High (A)	Low (B)	Indeterminate (C)		
	Grande (A)	Faible (B)	Indéterminée (C)		
Census omissions - Recensement					
Single omission - Omission unique:					
Number of records - Nombre de dossiers	3,842	2,139	2,640		8,621
Census total income(1) aggregate - Revenu agrégatif total au recensement(1)	\$'000 23,010	10,640	20,586		54,236
RC-T total income aggregate - Revenu agrégatif total à RC-I	\$'000 23,121	11,677	21,385		56,183
Non-sampling effect (NSE) - Erreur d'observation (EO)	\$'000 111	1,037	799		1,947
NSE/RC-T - Total income - EO/RC-I revenu total	% 0.5	8.9	3.7		3
Average NSE - EO moyenne	\$ 29	485	303		2
Two or three omissions - Deux ou trois omissions:					
Number of records - Nombre de dossiers	335	549	383		1,267
Census total income(1) aggregate - Revenu agrégatif total au recensement(1)	\$'000 2,320	2,765	3,715		8,800
RC-T total income aggregate - Revenu agrégatif total à RC-I	\$'000 2,334	3,798	3,816		9,948
Non-sampling effect (NSE) - Erreur d'observation (EO)	\$'000 14	1,033	101		1,148
NSE/RC-T - Total - EO/RC-I	% 0.6	27.2	2.6		11
Average NSE - EO moyenne	\$ 42	1,882	264		90
Revenue Canada omissions - Revenu Canada					
Single omission - Omission unique:					
Number of records - Nombre de dossiers	552	1,113	591		2,256
Census total income(1) aggregate - Revenu agrégatif total au recensement(1)	\$'000 3,169	9,555	4,260		16,984
RC-T total income aggregate - Revenu agrégatif total à RC-I	\$'000 3,156	4,784	4,079		12,019
Non-sampling effect (NSE) - Erreur d'observation (EO)	\$'000 13	4,771	181		4,965
NSE/RC-T - Total income - EO/RC-I revenu total	% 0.4	99.7	4.4		41
Average NSE - EO moyenne	\$ 24	4,287	306		2,200
Two or three omissions - Deux ou trois omissions:					
Number of records - Nombre de dossiers	6	140	32		178
Census total income(1) aggregate - Revenu agrégatif total au recensement(1)	\$'000 39	2,228	281		2,548
RC-T total income aggregate - Revenu agrégatif total à RC-I	\$'000 39	618	282		949
Non-sampling effect (NSE) - Erreur d'observation (EO)	\$'000 -	1,610	1		1,611
NSE/RC-T - Total - EO/RC-I	% -	260.5	0.4		171
Average NSE - EO moyenne	\$ -	11,500	31		9,000

(1) Made conceptually compatible; i.e., "income subject to taxation".

(1) Rendu conceptuellement compatible; c.-à-d., "revenu soumis à l'impôt".

TABLE 13. Income Effect of Component Substitution, by Reliability Category and Incidence of Substitution for Income Base Year, 1970

TABLERAU 13. Effet sur le revenu de la substitution d'un élément, par catégorie de fiabilité, selon la source de la substitution et le groupe d'incidence, 1970

Incidence group Groupe d'incidence	Reliability category - Catégorie de fiabilité			
	High (A) Grande (A)	Low (B) Faible (B)	Indeterminate (C) Indéterminée (C)	Total
Single component substitution - Substitution unique:				
Number of records - Nombre de dossiers	850	906	401	2,157
Census total income(1) aggregate - Revenu agrégatif total au recensement(1)	\$'000 4,968	6,523	3,090	14,581
RC-T total income aggregate - Revenu agrégatif total à RC-I	\$'000 4,970	3,920	3,098	11,988
Non-sampling effect (NSE) - Erreur d'observation (EO)	\$'000 2	2,603	8	2,593
NSE/RC-T - Total income - EO/RC-I revenu total	% --	66.4	0.3	21.6
Average NSE - EO moyenne	\$ 2(2)	2,873	20	1,202
Two or three component substitutions - Deux ou trois substitutions:				
Number of records - Nombre de dossiers	13	27	8	48
Census total income(1) aggregate - Revenu agrégatif total au recensement(1)	\$'000 86	292	44	422
RC-T total income aggregate - Revenu agrégatif total à RC-I	\$'000 86	194	47	327
Non-sampling effect (NSE) - Erreur d'observation (EO)	\$'000 -	98	3	95
NSE/RC-T - Total income - EO/RC-I revenu total	% -	50.5	6.4	29.1
Average NSE - EO moyenne	\$ -	3,630	375	1,979

(1) Made conceptually compatible; i.e., "income subject to taxation".

(1) Rendu conceptuellement compatible; c.-à-d., "revenu soumis à l'impôt".

(2) Within Census reporting (rounding) limits of \$10.

(2) Les chiffres du recensement ont été arrondis à \$10.

TABLE 14. Income Effect of Combined Omissions and Substitutions of Income Components, by Reliability Category and Source of Omissions for Income Base Year, 1971

TABLEAU 14. Effet sur le revenu des omissions et des substitutions des éléments du revenu, par catégorie de fiabilité et source des omissions, 1971

Source of omission Source d'omission	Reliability category - Catégorie de fiabilité			
	High (A) Grande (A)	Low (B) Faible (B)	Indeterminate (C) Indéterminée (C)	Total
Census - Recensement:				
Records with one or two omissions combined with one or two substitutions - Dossiers comportant une ou deux omissions et une ou deux substitutions	115	321	140	576
Census total income(1) aggregate - Revenu agrégatif total au recensement(1)	\$'000 734	2,150	1,226	4,110
RC-T total income aggregate - Revenu agrégatif total à RC-I	\$'000 738	1,979	1,235	3,952
Non-sampling effect (NSE) - Erreur d'observation (EO)	\$'000 4	171	9	158
NSE/RC-T - Total Income - EO/RC-I revenu total	% 0.5	8.6	0.7	4.0
Average NSE - EO moyenne	\$ 35	533	64	274
RC-T - RC-I:				
Records with one or two omissions combined with one or two substitutions - Dossiers comportant une ou deux omissions et une ou deux substitutions	13	112	33	158
Census total income(1) aggregate - Revenu agrégatif total au recensement(1)	\$'000 89	970	293	1,352
RC-T total income aggregate - Revenu agrégatif total à RC-I	\$'000 88	549	275	912
Non-sampling effect (NSE) - Erreur d'observation (EO)	\$'000 1	421	18	440
NSE/RC-T - Total income - EO/RC-I revenu total	% 1.1	76.7	6.5	48.2
Average NSE - EO moyenne	\$ 77	3,759	545	2,785

(1) Made conceptually compatible; i.e., "income subject to taxation".

(1) Rendu conceptuellement compatible; c.-à-d., "revenu soumis à l'impôt".

TABLE 15. Match Rates and Taxfiler Rates with Components, by Province with Descending Rank Order for the Income Base Year, 1970

TABLEAU 15. Éléments des taux d'appariement et des taux de déclaration à l'impôt, par province et par ordre décroissant, 1970

Province	Match rate component			
	Élément du taux d'appariement			
	True match	False non-match	Estimated tax universe in sample	Match rate(1)
	Appariement juste	Non-appariement erroné	Univers fiscal estimatif de l'échantillon	Taux d'appariement(1)
Newfoundland - Terre-Neuve	1,942	204	2,146	90.5
Prince Edward Island - Île-du-Prince-Édouard	510	64	574	88.9
Nova Scotia - Nouvelle-Écosse	3,814	356	4,170	91.5
New Brunswick - Nouveau-Brunswick	3,001	299	3,300	90.9
Québec	8,112	1,167	9,279	87.4
Ontario	12,647	901	13,548	93.3
Manitoba	2,434	206	2,640	92.2
Alberta	2,030	131	2,161	93.9
British Columbia(2) - Colombie-Britannique(2)	4,056	284	4,340	93.5
	5,340	472	5,812	91.9
TOTAL	43,886	4,084	47,970	91.5
	Taxfiler rate component			Descending rank order
	Éléments du taux de déclaration à l'impôt			Ordre décroissant
	Tax return(3)	Adult population(4)	Taxfiler rate(5)	Taxfiler rate
	Déclaration d'impôt(3)	Population adulte(4)	Taux de déclaration à l'impôt(5)	Taux d'appariement
Newfoundland - Terre-Neuve	153,131	327,520	46.8	8
Prince Edward Island - Île-du-Prince-Édouard	37,046	76,235	48.6	9
Nova Scotia - Nouvelle-Écosse	296,835	548,195	54.1	6
New Brunswick - Nouveau-Brunswick	231,151	431,455	53.6	7
Québec	2,307,452	4,242,225	54.4	10
Ontario	3,640,362	5,494,615	66.3	3
Manitoba	427,987	701,450	61.0	4
Alberta	357,963	645,815	55.4	1
British Columbia(2) - Colombie-Britannique(2)	690,442	1,113,370	62.0	2
	1,000,255	1,606,525	62.3	5
TOTAL	9,142,624	15,187,405	60.2	...

1) The percentage of true matches within the estimated tax universe.

1) Pourcentage des appariements justes au sein de l'univers fiscal estimatif.

2) Includes Yukon and Northwest Territories.

2) Yukon et Territoires du Nord-Ouest compris.

3) Filed early in 1971 for the 1970 taxation year. Source: Taxation Statistics (Revenue Canada - Taxation).

3) Produites au début de 1971 pour l'année fiscale 1970. Source: Statistiques fiscales (Revenu Canada - Impôt).

4) All persons 15 years and over on Census Day, 1971. Source: Census of Population, Catalogue 92-717, pp. 19-1 to 19-15.

4) Ensemble des personnes de 15 ans et plus le jour du recensement (1971). Source: Recensement de la population, n° 92-717 au catalogue, pp. 19-1 à 19-15.

5) Percentage of taxfilers within adult population.

5) Pourcentage de contribuables au sein de la population adulte.

TABLE 16. Substitution of Employment Income Components(1) for the Income Base Year, 1970

TABLEAU 16. Substitution d'éléments du revenu de l'emploi(1), 1970

	Component reported to RC-T but not to Census (substitution)			Supplementary components(2) (no substitution)		
	Éléments déclarés à RC-I, mais non au recensement (substitution)			Éléments supplémentaires(2) (aucune substitution)		
	Farm net income	Non-farm self-employment income	Wages and salaries	Farm net income	Non-farm self-employment income	Wages and salaries
	Revenu agricole net	Revenu non agricole, travail autonome	Rémunérations	Revenu agricole net	Revenu non agricole, travail autonome	Rémunérations
Component reported to Census but not to RC-T (substitution) - Éléments déclarés au recensement, mais non à RC-I (substitution):						
Farm net income - Revenu agricole net	...	54	56	...	235	251
Non-farm self-employment income - Revenu non agricole, travail autonome	105	...	433	1,309	...	735
Wages and salaries - Rémunérations	88	573	...	1,199	694	...
Supplementary components(3) (no substitution) - Éléments supplémentaires(3) (aucune substitution):						
Farm net income - Revenu agricole net	...	1,369	970
Non-farm self-employment income - Revenu non agricole, travail autonome	294	...	593
Wages and salaries - Rémunérations	315	694

(1) Cell entries are not mutually exclusive, but double counting is unlikely (see text).

(1) Les chiffres ne s'excluent pas mutuellement, mais les doubles comptes sont peu probables (voir texte).

(2) Components are absent in both sources or reported in both sources, given that initial (stub) component has been reported to Census; thus, no substitution.

(2) Éléments absents dans les deux sources ou déclarés dans les deux sources, l'élément initial (marge) ayant été déclaré au recensement; il n'y a donc pas substitution.

(3) Components are absent in both sources or reported in both sources, given that initial (heading) component has been reported to Revenue Canada - Taxation; thus, no substitution.

(3) Éléments absents dans les deux sources ou déclarés dans les deux sources, l'élément initial (titre) ayant été déclaré à Revenu Canada - Impôt; il n'y a donc pas substitution.

TABLE 17. Census Gains and Losses Vis-à-vis RC-T Reporting as a Result of Component Substitution for the Income Base Year, 1970

TABLEAU 17. Gains et pertes du recensement par rapport à RC-I résultant de la substitution d'éléments, 1970

Census income component	Gains(1)	Losses(2)	Net gain (+) net loss (-)
Éléments du revenu (recensement)	Gains(1)	Pertes(2)	Gains nets (+) pertes nettes (-)
Farm net income - Revenu agricole net	110	193	- 83
Non-farm net income from self-employment - Revenu non agricole net, travail autonome	538	627	- 89
Wages and salaries - Rémunérations	661	489	+ 172

(1) Sums of rows in upper left quadrant of Table 16.

(1) Somme des lignes du cadre supérieur gauche du tableau 16.

(2) Sums of columns in upper left quadrant of Table 16.

(2) Somme des colonnes du cadre supérieur gauche du tableau 16.

TABLE 18. Number and Percentage of Records Appearing in Equivalent and Neighbouring Income Classes for Selected Income Components whenever Component has been Reported in at Least One Source (Census or RC-T) for the Income Base Year, 1970

TABLEAU 18. Nombre et pourcentage de dossiers paraissant dans des classes de revenu équivalentes ou voisines en fonction de certains éléments du revenu déclarés dans au moins une source (recensement ou RC-I), 1970

Income type (number of classes compared) Genre de revenu (nombre de classes comparées)		Below main diagonal (Census over-reported)		On main diagonal (MD)
		Sous la diagonale principale (sur déclaration, recensement)		Sur la diagonale principale (DP)
		More than one class	One class	On MD or adjacent
		Plus d'une classe	Une classe	Sur la DP ou voisines
Total income(1) - (38) - Revenu total(1)				
	No. - nbre	5,655	3,734	24,734
	No. - nbre			(33,151)
	%	12.9	8.5	56.4
	%			(75.5)
Wages and salaries - (38) - Rémunérations				
	No. - nbre	4,359	3,213	23,375
	No. - nbre			(30,251)
	%	11.2	8.3	60.2
	%			(77.9)
Wages and salaries - (18) - Rémunérations				
	No. - nbre	2,930	2,916	26,061
	No. - nbre			(33,010)
	%	7.5	7.5	67.1
	%			(85.0)
Self-employment income - (24) - Revenu d'un travail autonome				
	No. - nbre	1,881	545	1,967
	No. - nbre			(2,943)
	%	30.6	8.9	32.0
	%			(47.9)
Non-employment income - (24) - Revenu hors-emploi				
	No. - nbre	2,592	1,209	8,600
	No. - nbre			(15,259)
	%	11.6	5.4	38.3
	%			(68.0)
		Above main diagonal (Census underreported)		Total compared
		Au-dessus de la diagonale principale (sous-déclaration, recensement)		Total comparé
		One class	More than one class	
		Une classe	Plus d'une classe	
Total income(1) - (38) - Revenu total(1)				
	No. - nbre	4,683	5,076	43,882
	%	10.7	11.6	100.0
Wages and salaries - (38) - Rémunérations				
	No. - nbre	3,663	4,244	38,854
	%	9.4	10.9	100.0
Wages and salaries - (18) - Rémunérations				
	No. - nbre	4,033	2,914	38,854
	%	10.4	7.5	100.0
Self-employment income - (24) - Revenu d'un travail autonome				
	No. - nbre	431	1,325	6,149
	%	7.0	21.5	100.0
Non-employment income - (24) - Revenu hors-emploi				
	No. - nbre	5,450	4,583	22,434
	%	24.3	20.4	100.0

1) Made conceptually compatible in both sources (income subject to taxation).

1) Rendu conceptuellement compatible dans les deux sources (revenu soumis à l'impôt).

TABLE 19. Distribution of Income Subject to Taxation with Class Deficiency Rates, by Match Status for the Income Base Year, 1970

TABLEAU 19. Répartition du revenu soumis à l'impôt et déficit, par statut d'appariement, 1970

Income class Catégorie de revenu	Match status Statut d'appariement				Match status (income source) Statut d'appariement (source de revenu)	
	True match Appariements justes	True non-match Non-appariements justes	False non-match Non-appariements erronés	Potential income universe Univers du revenu potentiel	True match (RC-T) Appariements justes (RC-I)	True non-match (Census) Non-appariements justes (recensement)
	number nombre				thousands of dollars milliers de dollars	
Loss - Écart	210	92	-	302	- 868	- 241
\$ 1-\$ 499	2,045	3,523	-	5,568	504	713
500- 999	2,364	2,314	-	4,678	1,792	1,808
1,000- 1,499	2,689	3,681	13	6,383	3,370	4,735
1,500- 1,999	2,676	967	211	3,854	4,683	1,597
2,000- 2,499	2,541	301	314	3,156	5,703	660
2,500- 2,999	2,400	150	221	2,771	6,605	404
3,000- 3,499	2,376	77	335	2,788	7,716	243
3,500- 3,999	2,401	41	295	2,737	9,001	152
4,000- 4,499	2,316	30	304	2,650	9,841	125
4,500- 4,999	2,167	16	227	2,410	10,286	75
5,000- 5,499	1,972	20	303	2,295	10,353	104
5,500- 5,999	1,905	4	162	2,071	10,940	23
6,000- 6,999	3,687	9	401	4,097	23,947	56
7,000- 7,999	3,070	16	404	3,490	22,966	121
8,000- 8,999	2,381	4	247	2,632	20,173	34
9,000- 9,999	1,819	4	191	2,014	17,231	38
10,000 and over - et plus	4,850	23	456	5,329	75,464	482
TOTAL	43,869	11,272	4,084	59,225	239,707	11,129
Match status (income source) Statut d'appariement (source de revenu)				Deficiency rate - Déficit		
				True non-match(1) Non-appariements justes(1)		False non-match(2) Non-appariements erronés(2)
				Records Dossiers	Income Revenu	Records Dossiers
				Income Revenu		Income Revenu
thousands of dollars milliers de dollars				per cent pourcentage		
Loss - Écart	-	- 1,109	30.5	21.7	-	-
\$ 1-\$ 499	-	1,217	63.3	58.6	-	-
500- 999	-	3,600	49.5	50.2	-	-
1,000- 1,499	17	8,122	57.7	58.3	0.2	0.2
1,500- 1,999	364	6,644	25.1	24.0	5.5	5.5
2,000- 2,499	689	7,052	9.5	9.4	9.9	9.8
2,500- 2,999	596	7,605	5.4	5.3	8.0	7.8
3,000- 3,499	1,064	9,023	2.8	2.7	12.0	11.8
3,500- 3,999	1,091	10,244	1.5	1.5	10.8	10.6
4,000- 4,499	1,258	11,224	1.1	1.1	11.5	11.2
4,500- 4,999	1,064	11,425	0.7	0.7	9.4	9.3
5,000- 5,499	1,559	12,016	0.9	0.9	13.2	13.0
6,000- 6,999	923	11,886	0.2	0.2	7.8	7.8
7,000- 7,999	2,564	26,567	0.2	0.2	9.8	9.7
8,000- 8,999	2,975	26,062	0.5	0.5	11.6	11.4
9,000- 9,999	2,060	22,267	0.2	0.2	9.4	9.3
10,000 and over - et plus	1,771	19,040	0.2	0.2	9.5	9.3
	7,039	82,985	0.4	0.6	8.6	8.5
TOTAL	25,034	275,870	19.0	4.0	6.9	9.1

(1) True non-matches or their aggregate income out of potential income universe.

(1) Non-appariements justes ou revenu agrégatif/univers du revenu potentiel.

(2) False non-matches or their aggregate income out of potential income universe.

(2) Non-appariements erronés ou revenu agrégatif/univers du revenu potentiel.

TABLE 20. Provincial Rank Order of Match Rates and Success Rates with Supporting Data for the Income Base Year, 1970
TABLEAU 20. Classement des taux d'appariement et des taux de réussite par province, 1970

Province	Adults in sample Adultes dans l'échantillon 1	False non-match Non-appariement erroné 2	Taxfiler rate(1) Taux de déclaration à l'impôt(1) 3	Estimated taxfilers in sample(2) Nombre estimatif de contribuables dans l'échantillon(2) 4	True match Appariement juste 5
Newfoundland - Terre-Neuve	4,464				
Prince Edward Island - Île-du-Prince-Édouard		204	46.8	2,089	1,942
Nova Scotia - Nouvelle-Écosse	1,013	64	48.6	492	
New Brunswick - Nouveau-Brunswick	7,263	356	54.1	3,929	510
Québec	5,863	299	53.6	3,143	3,814
Ontario	15,890	1,167	54.4	8,644	3,001
Manitoba	20,410	901	66.3	13,532	8,112
Saskatchewan	4,314	206	61.0	2,632	12,647
Alberta	3,849	131	55.4	2,132	2,434
British Columbia(3) - Colombie-Britannique(3)	6,884	284	62.0	4,268	2,030
	9,231	472	62.3	5,751	4,056
TOTAL	79,181	4,084	60.2	46,612(4)	5,340
					43,886
	Failure rate(5) Taux d'échec(5) 6	Success rate(6) Taux de réussite(6) 7	Match rate II(7) Taux d'appariement II(7) 8	Descending rank order Ordre décroissant	
				Success rate Taux de réussite 9	Match rate II Taux d'appariement II 10
					Match rate(1) Taux d'appariement(1) 11
Newfoundland - Terre-Neuve	4.6				
Prince Edward Island - Île-du-Prince-Édouard		95.4	93.0	4	8
Nova Scotia - Nouvelle-Écosse	6.3	93.7	103.7		8
New Brunswick - Nouveau-Brunswick	4.9	95.1	97.1	9	
Québec	5.1	94.9	95.5	6	9
Ontario	7.3	92.7	93.8	8	6
Manitoba	4.4	95.6	93.5	10	7
Saskatchewan	4.8	95.2	92.5	3	10
Alberta	3.4	96.6	95.2	5	3
British Columbia(3) - Colombie-Britannique(3)	4.1	95.9	95.0	1	4
				2	1
	5.1	94.9	92.9	7	2
TOTAL	5.2	94.8	94.1	...	9
				...	5
			

From Table 15.

Chiffres tirés du tableau 15

From Table 15.
Chiffres tirés du tableau 15.
Number of adults in sample multiplied by taxfiler rate.
Nombre d'adultes dans l'échantillon multiplié par le taux de déclaration à l'impôt.
Includes data from Yukon and Northwest Territories.
Yukon et Territoires du Nord-Ouest compris.
Estimated by adding provincial results in Column 4; when multiplying Column 1 by Column 3, 47,667 are estimated; the corresponding match rate II is 92.1%.
Somme des résultats de la colonne 4; si l'on multiplie la colonne 1 par la colonne 3, on obtient 47,667; le taux d'appariement II correspondant est donc de 92.1%.
False non-matches as a percentage of all adults in the sample.
Nombre de non-appariements erronés en pourcentage du nombre d'adultes dans l'échantillon.
One hundred per cent minus failure rate.
Cent pour cent moins le taux d'échec.
True matches as a percentage of Column 4 using taxfiler ratio as estimator. Estimates in excess of 100% are caused by overestimation of taxfiler population in Taxation Statistics, 1972. Thus match rate II is inferior to match rate in Table 15.
Appariements justes en pourcentage de la colonne 4, le pourcentage de contribuables étant utilisé comme estimateur. Les estimations supérieures à 100% sont attribuables à la surestimation de la population des contribuables dans Statistiques Fiscales, 1972. Le taux d'appariement II est donc inférieur au taux d'appariement du tableau 15.

BIBLIOGRAPHY

Dubois, Jr. N.S.D., "A Solution to the Problem of Linking Multi-variate Documents", **Journal of the American Statistical Association**, Vol. 64, 1969, pp.163-174.

Fellegi, I.P., and Sunter, A.B., "A Theory for Record Linkage", **Journal of the American Statistical Association**, Vol. 64, 1969, pp. 1183-1210.

Neter, J., Maynes, E.S., et al., "The Effect to Mismatching on the Measurement of Response Error", **Journal of the American Statistical Association**, Vol. 60, 1965, pp. 1005-1026.

Smith, M.E., and Newcombe, H.B., "Methods for Computer Linkage of Hospital Admission - Separation Records into Cumulative Health Histories", **Methods of Information in Medicine**, Vol. 14, 1975, pp. 118-125.

Tepping, B.J., "A Model for Optimum Linkage of Records", **Journal of the American Statistical Association**, Vol. 63, 1968, pp. 1321-1332.

U.S. Social Security Administration. "Some Observations on Linkage of Survey and Administrative Record Data", by Joseph Steinberg in **Studies from Interagency Data Linkages** (August 1973).

U.S. Social Security Administration. "Sub-sampling the Current Population Survey: 1963 Pilot Link Study", by Frederick Scheuren, Benjamin Bridges, and Beth Kilss in **Studies from Interagency Data Linkages** (Report No. 1, August 1973).

U.S. Social Security Administration. "Coverage Differences, Non-interview, Non-response, and the 1960 Census Undercount: 1963 Pilot Link Study", by Frederick Scheuren, Beth Kilss and H. Lock Oh in **Studies from Interagency Data Linkages** (Report No. 2, December 1973).

U.S. Social Security Administration. "Exact Match Research Using the March 1973 Current Population Survey - Initial Stages", by Frederick Scheuren, Roger Herriot, Linda Vogel, Denton Vaughan, Beth Kilss, Barbara Tyler, Cynthia Cobleigh, and Wendy Alvey in **Studies from Interagency Data Linkages** (Report No. 4, July 1975).

BIBLIOGRAPHIE

Dubois, Jr. N.S.D., "A Solution to the Problem of Linking Multi-variate Documents"; **Journal of the American Statistical Association**, vol. 64, 1969, pp. 163-174.

Fellegi, I.P. et Sunter, A.B., "A Theory for Record Linkage", **Journal of the American Statistical Association**, vol. 64, 1969, pp. 1183-1210.

Neter, J., Maynes, E.S., et al., "The Effect to Mismatching on the Measurement of Response Error", **Journal of the American Statistical Association**, vol. 60, 1965, pp. 1005-1026.

Smith, M.E. et Newcombe, H.B., "Methods for Computer Linkage of Hospital Admission - Separation Records into Cumulative Health Histories", **Methods of Information in Medicine**, vol. 14, 1975, pp. 118-125.

Tepping, B.J., "A Model for Optimum Linkage of Records", **Journal of the American Statistical Association**, vol. 63, 1968, pp. 1321-1332.

U.S. Social Security Administration. "Some Observations on Linkage of Survey and Administrative Record Data", Joseph Steinberg, **Studies from Interagency Data Linkages** (août 1973).

U.S. Social Security Administration. "Sub-sampling the Current Population Survey: 1963 Pilot Link Survey", Frederick Scheuren, Benjamin Bridges et Beth Kilss, **Studies from Interagency Data Linkages** (rapport n° 1, août 1973).

U.S. Social Security Administration. "Coverage Differences, Non-interview, Non-response, and the 1960 Census Undercount: 1963 Pilot Link Study", Frederick Scheuren, Beth Kilss et H. Lock Oh, **Studies from Interagency Data Linkages** (rapport n° 2, décembre 1973).

U.S. Social Security Administration. "Exact Match Research Using the March 1973 Current Population Survey - Initial Stages", Frederick Scheuren, Roger Herriot, Linda Vogel, Denton Vaughan, Beth Kilss, Barbara Tyler, Cynthia Cobleigh et Wendy Alvey, **Studies from Interagency Data Linkages** (rapport n° 4, juillet 1975).

Palmer, Gladys L., "Factors in Variability of Response in Enumeration Studies", **Journal of American Statistical Association**, June 1943, pp. 143-152.

Palmer, Gladys L., Factors in Variability of Response in Enumeration Studies", **Journal of American Statistical Association**, juin 1943, pp. 143-152.

Canadä

30/9/87

